Thèse de Doctorat

de

l'Institut Polytechnique de Paris

préparée à

Télécom Paris

en défense du titre de
Docteur en Mathématiques

Guillaume Ausset

à Paris-Saclay

Thèse présentée et soutenue à Paris-Saclay le 02.12.2021

Méthodes d'apprentissage statistique pour l'analyse prédictive du risque de crédit

par

Guillaume Ausset

Encadrants : Prof. Stephan Clemençon
Prof. François Portier
Dr. Timothée Papin

Jury : Prof. Marc Hoffmann
Prof. Anouar El Ghouch
Prof. Mathieu Rosenbaum
Prof. Cristina Butucea
Dr. Tabea Rebafka

—neque, me ut miretur turba, laboro:
Contentus paucis lectoribus

Colophon

This document was typeset using LuaLaTeX, using the KOMA-Script class and largely inspired by the typesetting of Edward Tufte's Beautiful Evidence, Robert Bringhurst's The Elements of Typographic Style. Large part of the code and layout have been borrowed from the thesis of Aaron Turon and Ken Arroyo Ohori. The bibliography is processed using BibTeX and biber. Slimbach's Minion Pro acts as both the text and display type-face with matching mathematical characters from Johannes Küster's Minion Math. Sans-serif text is set in Paul Renner's Futura, digitized by the Paratype foundry. Monospaced text is typeset with Fabrizio Schiavi's PragmataPro.

Contents

Figures

Tables

Remerciements

Rédiger les remerciements est un exercice difficile. Non seulement il s'agit, avec une grande probabilité, probablement de la seule partie qui sera lue dans son entièreté par beaucoup et donc de la partie la plus importante de la thèse ; mais surtout il est impossible de rédiger les remerciements tels qu'ils devraient être en toute logique. Cette thèse n'a été possible que grâce à l'aide et le soutien d'un grand nombre de personnes et remercier chacun à sa juste valeur nécessiterait plus de place que la thèse en elle-même. Il ne s'agit donc ici que d'un *abstract* de remerciements.

Je tiens tout d'abord à remercier mes encadrants François et Stéphan qui m'ont accompagné pendant presque 4 ans et qui sans leur pédagogie et patience rien n'aurait été possible. Je vous suis éternellement redevable et je suis fier d'avoir appris de vous. Merci à tous les lecteurs et le jury en premier, de me faire l'honneur de me lire. J'espère que cette modeste thèse vous apportera quelque chose.

Merci à Timothée et toute l'équipe du MSH : Sami, Mohamed, El Mostapha, Anne, Frédéric, Mourad, Olivier et tous les autres, qui m'ont accueilli et m'ont permis de réaliser ma thèse dans les meilleures conditions possibles au sein de la BNP et ce même sans avoir à faire d'Excel ou Visual Basic. Un merci tout particulier à l'équipe Quant qui m'a aidée dans tous les projets et a accepté de me suivre sur Python, Julia et Linux sans (trop) broncher.

Merci Arnaud de me supporter depuis Dieu seul sait combien de temps et d'avoir toujours été la pour me soutenir même dans les projets bizarres et même si cela veut dire racheter mon matériel informatique pour me permettre de financer d'*autres* projets que tu sais très bien n'aboutiront pas. D'ailleurs, tu ne voudrais pas me racheter une police d'écriture ?

Merci à toi, Tania, pour ton amitié sans faille et ton soutien permanent depuis maintenant 6 ans même si j'ai l'impression que c'était hier. C'est plus difficile de réussir les projets de mathématiques sans ton aide précieuse, j'espère que celui-là sera au niveau ! Ta volonté et ton sérieux m'ont inspiré dans les moments de fatigue. Merci pour tout, ma très chère amie, Спасибо тебе за все.

Merci, Claudia, d'avoir été là dans tous les moments difficiles y compris celui-là et de m'avoir écouté me plaindre en permanence sans jamais faillir, tu es une sainte. Difficile de croire que lorsque nous nous sommes rencontrés je finissais mes études, et je finis *encore* mes études. J'ai maintenant le temps pour un kebab alors attends-toi à ce que je fasse valoir mon bon pour un repas, avec intérêts !

Bien qu'un simple merci ne pourra jamais suffire pour exprimer ma gratitude et mon amour, merci à vous, Antoine, Papa et Maman. Si j'en suis arrivé là, c'est grâce à vous et la chance extraordinaire que j'ai eue de vous avoir. Merci pour tout ce que vous avez fait pour moi et maintenant

que j'ai enfin fini mes études je vais enfin pouvoir vous le rendre. Cette thèse est pour vous. Je vous aime à l'infini et retour.

Et enfin, merci à toi Paula. Merci d'avoir commencé une thèse et m'avoir donné envie d'en faire une. Je serais probablement encore en train de faire du Java sinon, et je ne souhaite ça à personne ! Merci pour ton soutien sans failles, toutes les heures, tous les jours, toutes les semaines, pendant 4 ans de thèse. Maintenant que c'est fini, il ne reste plus qu'à apprendre le Norvégien pour la suite. Te amo demasiado.

Merci à tous ceux qui ont été là pour moi au cours de ma vie et que jamais je n'oublie : Denys, François, Aleksander, Mariska, Adel, Laura, Lia, Katya, Anne, Clément, Thomas, Camille et tous les autres.

Merci à ceux qui ne sont plus là, mais à qui je dédie cette thèse. Je vous aime Papy et Mamy, vous me manquez, mais au final j'ai réussi, j'ai un diplôme de Polytechnique.

nulla dies umquam memori vos eximet aevo

Introduction 1

In this thesis we will study the death of individuals with the goal of predicting it from their characteristics. This task, known as *survival analysis* and originally intrinsically linked to that of epidemiology, has a rich mathematical history and have evolved to follow the successive advancements in statistics. Modelling the death of an individual have remained for centuries one of the flagship problems of medical research and biostatistics for the simple reason that understanding the cause of death is a first step to preventing said death. In parallel, statisticians have provided medical researchers with the mathematical tools necessary to answer the medical questions in properly scientific manner and to compare the survival of subpopulations as well as quantify the certainty of their hypothesis. This thesis will have plenty of such medical examples as it is not only an interesting and worthwhile endeavour, but because medical researchers have generously gifted the scientific community with a trove of open datasets. This memoir is, however, not about medicine but about finance and it is not about understanding causes and effects or statistically proving statements, but only about predicting death. This thesis, while in large part theoretical, had for main objective to answer the practical needs of BNP Paribas and in particular the Portfolio Management department of the CIB branch whose main role is to reduce the exposure of the bank to credit risk by actively managing this risk. In order to manage this risk we need to accurately predict the potential events, often using highly unstructured and voluminous data, which naturally motivates a rigorous study of credit risk through the scope of both survival analysis and machine learning.

1.1 Life and Death of a Company

We have mostly described the study of death as the study of death of individuals but it is not reserved to living beings. Even in everyday language, it is common to refer to the catastrophic failure of an item as its *death*; "My phone died!". If death can occur for inanimate objects too then it seems sensible, or at least useful, to predict it before it happens maybe to budget[1] for a new phone before the previous one dies. In this specific case, that

1: In a perfect world at least. Or more realistically if one is in charge of a fleet of thousands of corporate phones.

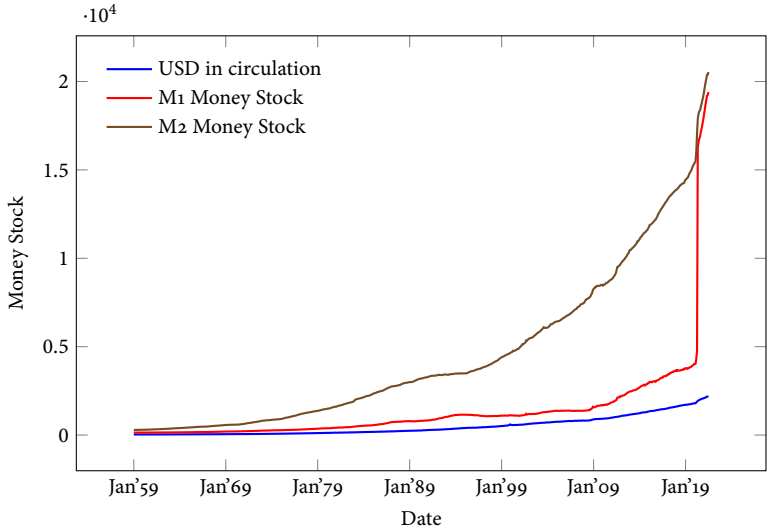


FIGURE 1.1: Physical money *vs* money stock.

is predicting the failure of a mechanical item, predicting failure before it happens makes it possible to schedule maintenance in advance in order to be parsimonious on expensive maintenance checks and operations. This problem, known in the literature as predictive maintenance (see Zonta et al. [2020]; Bousdekis et al. [2019]; Ran et al. [2019], for an overview), is naturally amenable to being treated as a survival analysis problem (see e.g. C. Chen et al. [2020]).[2] A similar interpretation naturally arises in the financial world through the concept of *credits*. Since the advent of banking and later fractional reserve, credits have come to represent the majority of the money-like assets in circulation[3] as it frees liquidity that can be used in the economy. This phenomenon has greatly accelerated in recent years as people have come to accept the dissociation between the concepts of money, i.e. a means of exchange, and physicality. In fig. 1.1 we represent the money stock of the US dollar, where the M1 stock encompasses currency outside the U.S. Treasury, deposits at commercial banks and other checkable deposits and the M2 stock consists of the M1 stock plus savings deposits and balances in retail money market funds. While not inherently a problem, the fact that most of the balance of companies or even individuals[4] now consists of loans which carry counterparty risk means that wealth has to be treated as a random variable. A *default* is the death of a loan.

2: It is also possible to see it as bandit problem (Ruiz-Hernández, Pinar-Pérez, and Delgado-Gómez [2020]; Fouché, Komiyama, and Böhm [2019]).

3: One could argue that money itself only represents an IOU or credit in physical form.

4: The money I have today in my bank account is a loan, which hopefully my bank will not default on.

1.1.1 Defaults and Contagions

While, from the point of view of the borrower, loans can effectively be considered discounted money as they receive cold hard cash, the same cannot be said at all for the lender. From the lender's point of view, loans carry a significant amount of uncertainty or risk called *counterparty* risk[5]: the borrower may very well never repay its loan. The value of a loan, that is the amount of money the loan will bring,[6] is therefore a random quantity. If the borrower repays its loan in entirety then the realized value will be the loan plus interest while if, for some reason, the loan is not repaid the value is only equal to the principal and interests repaid until default. Clearly, the realized gain is stochastic in nature and depends on the random event "repaid its loan" of which the probability is primordial. From this observation it is possible to define the *fair value* of a loan, which for simplicity we will here assume to be the expected profit,[7] and from this definition of fair value it is possible to find the rate at which a loan should be emitted. There is, however, one significant drawback to the previous remark: by reasoning in terms of expected value we hide the fact that individuals only have a finite amount of money and therefore can only survive a finite amount of losses. While this would not be a problem if all entities had more cash on hand than outstanding loans, we have seen earlier in fig. 1.1 that, for good reasons, the amount of money tied in loans vastly outsize the amount of money held. It is therefore possible, and even guaranteed given enough time (see Embrechts, Klüppelberg, and Mikosch [1997], for ruin theory, e.g. fig. 1.2), that an extreme event[8] occurs that is greater than the lender can bear. While such catastrophic events are in theory rare, their impact can have catastrophic repercussions for the same reasons viruses do. Individuals, or in this case companies, do not exist in a vacuum and interact with each other. While with the plague or the smallpox this interaction can result in spreading a viral load, in the case of companies the interaction results in spreading credit losses. A company suffering an extreme credit loss resulting in its own default is an incredibly rare event taken individually, but a systematic failure of a multitude of loans and companies is a much more probable event conditioned on this primary failure, or *patient zero*. A company suffering a loss sufficient to provoke its own default is, by definition, not capable of honouring its own loan which depending on the size of the entity collapsing can provoke the subsequent default of other companies. As more companies default on their loan, the phenomenon propagates through the intricate web of financial relations and can in the worst case provoke a financial crisis. Such events led to the mortgage crisis of 2008 where the correlation between entities and therefore the risk of spread of this financial virus was underestimated.[9] Given the parallels between biological viruses and credit risk, it should

5: In the most general setting. We do not differentiate here, for the sake of simplicity, between the credit risk and counterparty risk.

6: Eventually discounted at the risk free-rate, but we will ignore all the purely financial concerns here.

7: In practice all financial products are valued as the *expected value* of their payoff, but the expected value is not necessarily under the natural probability but often under a different probability called the risk-neutral measure. This is out of scope of this thesis but curious readers can refer to Shreve (2004).

8: That is, en event deep in the tail of distribution of losses.

9: Readers of this thesis may also have the more recent Evergrande's case as example depending on how events unfolded.

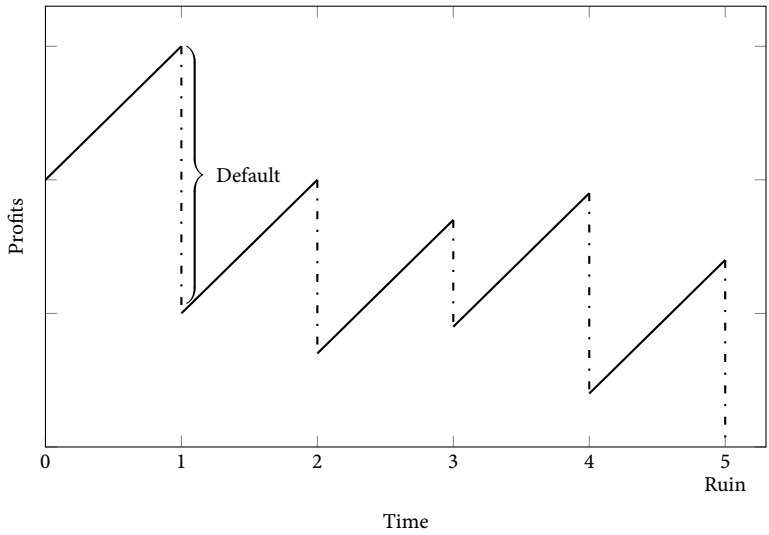


FIGURE 1.2: Assets of a company accrued from loans seen as a compound Poisson process.

not be surprising that people looked toward pandemics for inspiration of treatments of financial crisis.

1.1.2 Regulatory Vaccination

After the crisis of 2008, unprecedented[10] measures were implemented in order to *vaccinate* the financial world against credit risk. While vaccines rely on giving a base minimum amount of antigens in order to survive a normal viral load,[11] credit regulations such as BASEL II to BASEL IV (Basel Committee [2019]), rely on ensuring a minimum amount of cash buffer to survive extraordinary, or "unexpected"[12] as represented in fig. 1.3, losses without defaulting and therefore without contaminating other counterparties. In order to ensure resiliency of financial actors to extreme losses, those are required to hold a minimum amount of capital to compensate for the risky assets denoted here by RWA such that

$$\frac{\text{Capital}}{\text{RWA}} \geq 0.08.$$

In order to harmonize approaches, and more importantly prevent dubious calculations for the risk weighted assets (RWA) and therefore severe underestimations[13] of the capital required, the Bank for International Settlements (BIS) provides a common way of determining the quantities required. For

10: In size, the general idea and framework existed beforehand.

11: Please, do not use my explanations as anything more than a metaphore, see A. J. Pollard and Bijker (2021).

12: Unexpected in the business sense, not the mathematical sense. Abuse of mathematical terms is a longstanding tradition in the financial world next to that of inventing imaginary greek letters.

13: Voluntarily or not.

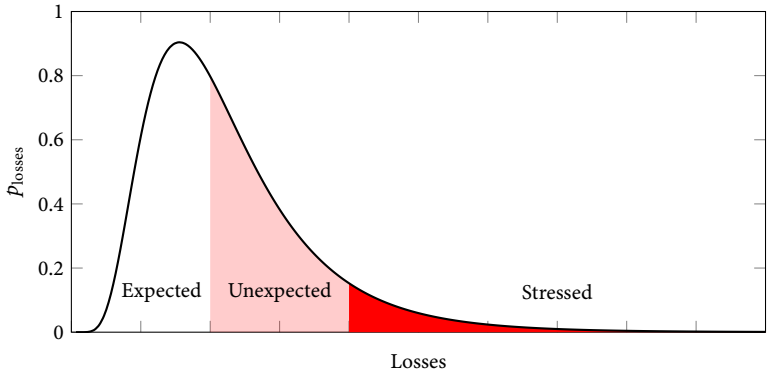


FIGURE 1.3: Expected, unexpected and stress losses.

corporate loans for example, the Basel III framework imposes

$$\text{Capital} = \text{LGD} \times \left(\Phi\left(\sqrt{\frac{1}{1-R}}\Phi^{-1}(p) + \sqrt{\frac{R}{1-R}}\Phi^{-1}(0.999) \right) - p \right),$$

$$\text{RWA} = \frac{\text{Capital} \times \text{EAD}}{0.08},$$

where R is a correlation factor defined by

$$R = A\left(0.12 \times \frac{1-e^{-50p}}{1-e^{-50}} + 0.24 \times \left(1 - \frac{1-e^{-50p}}{1-e^{-50}}\right)\right),$$

and $A \in \{1, 1.25\}$ is a factor depending on the size of the institution,[14] Φ is the cumulative distribution function (CDF) of the standard normal and p is the probability of default. As the probability of default is the only quantity not explicitly given, and the driver behind the previous values as quickly illustrated in fig. 1.4, its estimation plays a central role in the business strategy of financial organizations. Indeed, while there is no serious drawback to vaccination in humans other than a potentially sore arm and flu-like syndrome, this is not the case for financial actors. Preparing oneself against catastrophic default and following the regulations entails freezing a significant amount of money in risk-free assets[15] and therefore at a significant opportunity cost and at worst real economic losses. While the regulator provides guidelines on the estimation of p based on the rating given by external agencies, it also offers some liberties and allow this key quantity to be modelled internally using the so-called advanced internal ratings-based (A-IRB) approach.

Most approaches to credit risk treat the problem as a classification problem, predicting either default or no-default. However this approach has serious drawbacks as it relies on discretizing time and deciding arbitrarily to classify something as a default or non-default based on a single time

14: To represent how central and connected to others the institution is.

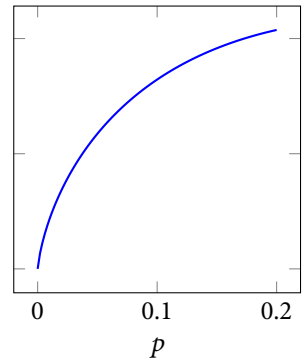


FIGURE 1.4: Capital required depending on the probability of default.

15: Returns in finance are directly related to risk. Risk-free assets therefore offer no returns or even negative returns.

horizon τ. In this approach, a company defaulting after $\tau + 1$ days is considered as a good payer, a very arguable design decision. For this reason, and because the exact same problems are encountered in the medical setting, we adopt a different approach: instead of predicting a binary event, "default" *vs* "non-default"[16] we predict the *time until default*, or more generally *time-to-event*, by adopting the point of view that all companies end up defaulting and we only potentially have not been able to observe it.

16: Or "death" *vs* "no-death" or "failure" *vs* "no-failure".

FIGURE 1.5: *Memento mori*. For Benjamin Franklin, "Nothing can be said to be certain, except death and taxes". We ask the reader to add defaults for the remainder of this thesis.

1.2 Censored Time-to-Event

Predicting if an event happens can be tackled from multiple perspectives of which the simplest is to treat the problem as a binary classification problem. After choosing a temporal threshold τ it is possible to reframe most problems involving the survival of an individual, living or not, as the classification problem "did the event happen before τ or not?". This simplistic approach can be well adjusted to many problems where the act of choosing a threshold is in itself natural, for example for a specific loan with a predetermined duration τ, but is often wildly inadequate in practice. Most problems cannot be naturally thresholded; for a perpetual credit line if one fix the threshold at τ does it mean that clients defaulting at $\tau + 1$ are good clients? In the medical setting, if the goal is to compare two treatments where should τ be fixed? For a large τ you potentially only observe natural deaths or even nothing at all if the study is too short and for small a τ you incur the risk of not having waited long enough to observe anything. Moreover, by treating the problem as a classification task, we are quickly confronted to problems of imbalances, as luckily most clients do not default and most people do not die, which results in hard classification instances.

For all these reasons and more, it is therefore more natural to treat the problem of predicting an event as the problem of predicting a *time-to-event*, that is learning the distribution of the times at which the event of interest happens. If we denote the time-to-event by $Y \in \mathbb{R}_+$, following the usual notation of regression, which we suppose by convention to be a positive random variable, we then wish to estimate the distribution of Y through one of the multiple objects that define it. As the subject is called *survival analysis* and we are interested in the deaths, failures and defaults and not in the period of time where nothing happens, we usually chose to study the survival function $S(t) = \mathbb{P}(Y > t)$ instead of the CDF $F(t) = 1 - S(t)$. The survival function plays here the role of natural extension of the previous approach by thresholding as each instance $S(\tau)$ represents a binary classification problem, we therefore here solve *all* such instances at the same time. Of course other quantities of interest can be

modelled given the particularities of the problem as described later in fig. 3.1. In particular in the case where Y admits a density $p(t)$,[17] we can define the instantaneous hazard

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{\mathbb{P}\left(Y \in [t, t+\Delta t] \mid Y > t\right)}{\Delta t},$$

which relates to the survival naturally through the relation

$$\lambda(t) = \frac{p(t)}{S(t)}.$$

Similarly, the integrated or cumulative hazard $\Lambda(t) = \int_0^t \lambda(t)\,\mathrm{d}t$ is often studied because of the relation

$$S(t) = \exp(-\Lambda(t)),$$

which trivially derives from the definition of λ. Note that all those quantities uniquely define the law of Y and can therefore be used interchangeably.

While the time-to-event formulation is particularly well suited to our problem, we still have unfortunately have to deal with a time horizon. Not only are our observations necessarily stopped at the current moment, or at least when we stopped gathering data,[18] but some observations are unobserved for reasons beyond our control such as a patient dropping from a study or a company merging with another and effectively "disappearing". In reality we therefore can rarely observe Y, the real variable of interest, and instead only observe some time T which we call *right censored* such that

$$T = \min(Y, C)$$

where $C \in \mathbb{R}_+$ is a nuisance random variable, playing a symmetrical role to Y and here called the *censoring variable*, which encompasses all the reasons for which our variable of interest can be unobserved such as the arrow of time, the end of the study, an observation being taken out of the dataset etc. We also assume that we know whether the time we observe is censored or not through the *censoring indicator* δ defined by

$$\delta = \mathbb{1}_{Y \leq C} = \begin{cases} 1 & \text{if } Y \leq C \\ 0 & \text{otherwise.} \end{cases},$$

as otherwise there would be no hope of estimating any useful quantity. Our dataset, represented in fig. 1.6, therefore consists of observations of the tuple (T, δ) instead of Y.

This setting, the flagship setting of survival analysis,[19] as been extensively studied in the statistical literature (see Fleming and Harrington [1991]; or Gill [1994], for an excellent overview of the methods involved to derive estimators) which has focused on the asymptotic properties of the various estimators of S, Λ.

17: The unusual notation p instead of the more common f for the density will be used in this thesis.

18: For example when a medical study ends.

19: It is possible to also define left censoring as well as left/right truncation. Most of the results presented here can be straighforwardly, if not with growing technical pain, be adapted to the more general cases.

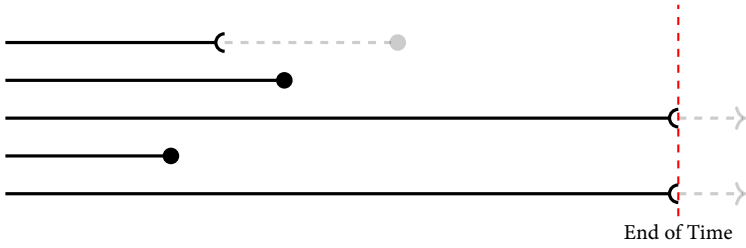


FIGURE 1.6: Right censored survival data.

1.2.1 Estimators of the Survival

As the field of survival analysis is too vast to summarize here, we will only introduce briefly the key estimator of the survival that will be the inspiration for later chapters. A rigorous survey of the literature is deferred to the relevant chapters. However, if the reader is interested in survival analysis, we recommend Klein and Moeschberger (2003); D. R. Cox and Oakes (1984) for a general overview as well as the previously mentioned course of Gill (1994) which motivates the product-integral formulation.

In the case where the object of interest is the survival S,[20] had we observed the real variable of interest Y we could easily estimate S by

$$S_n(t) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{Y_i>t},$$

which by Glivenko-Cantelli converges uniformly to the true survival. Here we only observe (T, δ) and the corresponding estimator would be

$$\bar{S}_n(t) = \frac{\sum_{i=1}^{n} \mathbb{1}_{T_i>t,\delta_i=1}}{\sum_{i=1}^{n} \mathbb{1}_{\delta_i=1}},$$

which is biased and do not converge to the value of interest.

However, after discretizing the time at each observation T_i, we can apply Bayes formula after noticing that locally, inside an interval $[T_{[i]}, T_{[i+1]}]$ where $T_{[k]}$ is used to mean "the k-th largest value of (T_i)" (when ignoring ties for simplicity), we can write the conditional probability of an event happening in this interval given that nothing happened until now as if no censoring was present. That is, if we denote by m_i the number of events in the i-th interval, n_i the number of *at risk individuals*, that is alive and non-censored and c_i the number of censored individuals at the beginning of the interval i.e. at T_i then

$$\mathbb{P}\left(Y \leq T_{[i+1]} \mid Y > T_{[i]}\right) = \frac{m_i}{n_i - c_i}.$$

20: Which often is the case. In the medical setting we are for example interested in comparing S_1 to S_2, the survival functions of two competing treatments, in order to prove some hypothesis of the type $S_1 > S_2$ or at least $S_1 \neq S_2$

We can therefore construct iteratively an estimator of S of the form

$$\hat{S}_n(t) = \prod_{i|T_i \leq t} \left(1 - \frac{m_i}{n_i - c_i}\right),$$

or rewritten under several different equivalent forms

$$\begin{aligned}\hat{S}_n(t) &= \prod_{i=1}^{n} \left(1 - \frac{\delta_{[i]}}{n-i+1}\right)^{\mathbb{1}_{T_{[i]} \leq t}} \\ &= \prod_{\substack{i=1,\dots,n \\ T_{[i]} \leq t}} \left(\frac{n-i}{n-i+1}\right)^{\delta_{[i]}}.\end{aligned}$$

This estimator, often called the Kaplan-Meier estimator (Kaplan and Meier [1958]), can be shown to be consistent. Similar results can be obtained for the cumulative hazard Λ with the Nelson-Aalen estimator (Nelson [1969]; Aalen [1978])

$$\hat{\Lambda}_n(t) = \sum_{i=i}^{n} \frac{\delta_i \mathbb{1}_{T_i \leq t}}{\sum_{j=1}^{n} \mathbb{1}_{T_j > T_i}}.$$

While we have here ignored the covariates X and therefore the conditioning on X, those estimators are of particular interest to us because of the ease of introducing this conditioning: by introducing local averaging around X, by example through kernels, we can obtain conditional estimators at $X = x$ of the form

$$\tilde{\Lambda}_n(t \mid X = x) = \sum_{i=i}^{n} \frac{\delta_i \mathbb{1}_{T_i \leq t} K(x - X_i)}{\sum_{j=1}^{n} \mathbb{1}_{T_j > T_i} K(x - X_j)},$$

where K is usually a probability density function[21] symmetric around 0.

1.2.2 Parametric and Semi-Parametric Models

21: Or kernel.

A surprising observation which unlocks a vast number of techniques available in the uncensored setting lies in the conditional decomposition of the likelihood of the observations. If we assume for a moment that Y and C are independent, then we can write the likelihood of the observation i as

$$\begin{aligned}&\mathbb{P}\left(T \in [T_i, T_i + \mathrm{d}t], \delta = \delta_i \mid \theta\right) \\ &\quad = \mathbb{P}\left(T \in [T_i, T_i + \mathrm{d}t], \delta = 1 \mid \theta\right)^{\delta_i} \\ &\qquad \times \mathbb{P}\left(T \in [T_i, T_i + \mathrm{d}t], \delta = 0 \mid \theta\right)^{1-\delta_i} \\ &\quad = \mathbb{P}\left(Y \in [T_i, T_i + \mathrm{d}t], C \geq T \mid \theta\right)^{\delta_i} \\ &\qquad \times \mathbb{P}\left(C \in [T_i, T_i + \mathrm{d}t], C < T \mid \theta\right)^{1-\delta_i} \\ &\quad = \left(p\left(T_i \mid \theta\right) S_C(T_i-)\right)^{\delta_i} \left(p_C(T_i) S(T_i \mid \theta)\right)^{1-\delta_i}, \qquad (1.1)\end{aligned}$$

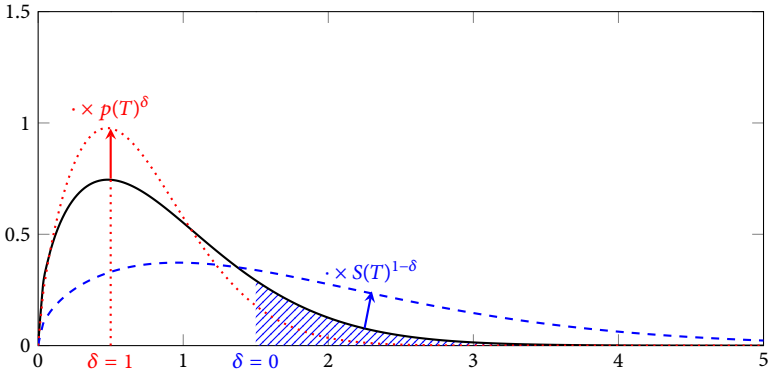


FIGURE 1.7: Individual Contribution of Censored and Uncensored Observations.

where $\theta \in \Theta$ is here the parameter describing the family of interest, that is the distribution of the survival and p_C, S_C are the density and survival of the censoring variable. The previous quantity in eq. (1.1) involves both the objects of interest p and S but also the nuisance quantities p_C, S_C which seems at first glance a problem. However as C is precisely a *nuisance* variable and therefore irrelevant to us; it is not useful to model it and it therefore does not involve θ in any way. As such, after ignoring those quantities by treating them as constants, we can write the likelihood as

$$\mathcal{L} \propto \prod_{i=1}^{n} p(T_i \mid \theta)^{\delta_i} S(T_i \mid \theta)^{1-\delta_i}. \tag{1.2}$$

The quantity in eq. (1.2) is often called the *partial likelihood*, and can be directly used for maximum likelihood estimation as the hidden constant that absorbed the nuisance quantities do not modify the solution in any way. The previous derivation of the partial likelihood is a fairly natural consequence of the very specific structure of the right censored problem of survival analysis as can be seen in fig. 1.7; as it simply expresses that when the observation is the true quantity we can update our knowledge in the usual way meanwhile when the observation is censored the most we can learn is "the true time-to-event is greater than the current observation".

As eq. (1.2) makes maximum likelihood estimation possible it is possible to approach the survival analysis problem through parametric modelling, by taking care of choosing a parametric family with support in $\mathbb{R}_+$ which is for example the approach we will use in chapter 3. Because of its intimate link to medical research, it is common in survival analysis not to be interested by S itself but by the comparison of S_1 and S_2, the cause specific survivals of two populations, for example corresponding to a reference or placebo treatment and a new treatment or more generally of $S(\cdot \mid X_1)$ compared to $S(\cdot \mid X_2)$. In this case, semi-parametric approaches have

enjoyed great success of which the proportional hazard model of D. R. Cox (1972) is certainly the most iconic representant. In the proportional hazard model, often referred to as Cox's model; the hazards are supposed to be proportional such that

$$\lambda(t \mid X) = \lambda_0(t) \exp\left(\theta^\intercal X\right),$$

where λ_0 is intentionally kept as nonparametric and entirely general. The partial likelihood can then be written as

$$\sum_{i:\delta_i=1} \left(\theta^\intercal X_i - \log \sum_{j:T_j \geq T_i} \exp\left(\theta X_j\right) \right),$$

which surprisingly do not involve λ_0 in any way and can therefore by learned with ease. Note, however, that if $\lambda(\cdot \mid X)$ is of interest and not solely $\lambda(\cdot \mid X_1)/\lambda(\cdot \mid X_2)$, then it is actually possible to estimate λ_0 nonparametrically (Breslow [1975]).

Similarly, there is a rich literature on regression models such as the accelerated failure time (AFT) model (Buckley and James [1979]) where Y is modelled as

$$\log(Y) = -\log\left(f(X)\right) + \epsilon,$$

with ϵ a baseline distribution, or Poisson regression.[22] Those many methods, however, have significant flaws we would like to alleviate. Most of the results from the statistical literature make the assumption that the estimated model and the true generative model are in the same class which is, of course, never true but was often accepted as unavoidable in order to derive interesting theoretical results. We would, however, prefer results that match the reality of the data; that is, bounds that do not assume the model to be correct, even if the resulting bounds are necessarily less tight. Similarly, most results deal with convergence, and its characterization, in the asymptotic regime. While the results obtained are very powerful, those are of little utility to practitioners confronted with finite samples. Those last two remarks form the basis of *statistical learning theory*, or what most people have come to refer to as *machine learning*. This work therefore tries to bridge the world of survival analysis with that of machine learning in order to bring the type of theoretical guarantees practitioners in machine learning have come to expect, to the existing tools of survival analysis.

22: The Kaplan-Meier estimator of the previous paragraph can actually be derived by non-parametric maximum likelihood estimation too.

1.3 Censored Prediction in High-Dimension

As mentioned in passing in the previous section, survival analysis as a field is mostly born out of necessity to describe and understand natural phenomenons. The small historical presentation of the field given in

appendix B gives several such examples of how models can be used to help build a better understanding of the world in order to take decisions. This approach to modelling is certainly the most natural to most as it is both the historical one but more importantly the one people have been exposed to in their life. *Describing* nature is, of course, of the foremost importance to epidemiologists, virologists, econometrists or any other scientific field but generally industrial practitioners are often satisfied with much simpler but practical results. When trying to shoot a basketball through a hoop it is certainly useful to understand mechanics from first principle in order to know that the ball will follow a parabola, but it is more than sufficient to just predict that the ball will just go *that way* if you throw it *this hard* without understanding anything about Newton's laws of motion. The same principle applies to many analytical fields and in our case medicine and finance. If the goal is simply to predict a time-to-event, and not to understand the reasons behind this event, then it is sufficient to adopt a *predictive* point of view. We call here *prediction* the task of guessing, that is building an estimator of some quantity Y from the input or characteristics X on the hope that $Y = f(X)$. For our basketball player, X is the angle and force of the launch while in the medical setting, X would be the characteristics of the patient. In this point of view, the function f is some abstracted black box encompassing all the dynamics leading to the result as only the result $Y = f(X)$ is of interest. We have previously given the example of the Cox regression model, where the survival of some individual is modelled through the instantaneous hazard rate λ such that

$$\lambda(t \mid X) = \lambda_0(t) \exp(\beta^\intercal X) .$$

While this model can and is often used as a predictive model, its primary reason of being is the study of the relative impact of the different variables through the study of the coefficients β_i; the goal is to *understand* the mechanisms leading to death. On the other hand, if the goal is only to guess the quantity $Y = f(X)$ then it is sufficient and simpler[23] to decide on some criterion of the goodness of fit $\mathcal{L}$ in order to try to find the best possible f given the data for this criterion. Mathematically, we can express this vague goal as solving

$$\operatorname*{argmin}_f \mathbb{E}\left[\mathcal{L}\left(Y, f(X)\right)\right], \tag{1.3}$$

which is often referred to as the *risk minimization* problem. This formulation, while at first glance unnatural, actually encompasses many of the usual questions about the data one can have depending on the choice of $\mathcal{L}$. For example, taking $\mathcal{L}$ to be the squared loss $(Y - f(X))^2$ leads to the solution of eq. (1.3) being the conditional expectation $\mathbb{E}[Y \mid X]$, while the absolute value $|Y - f(X)|$ leads to the median. Similarly, quantities like the

23: We use here *simpler* to mean that we can expect in practice the results to be better on this task using the same data, compared to first estimating the distribution and then forming a plugin regressor.

quantiles or the conditional probability can also be obtained in a similar manner by choosing appropriate losses $\mathcal{L}$[24] and the *art* of choosing the correct loss for the task of interest attracts considerable research attention. We emphasize that *estimation* i.e. learning the conditional density or even conditional expectation is not the goal pursued here, and is instead only *prediction* by learning a predictive rule f with good generalization properties. While the same objects may be involved in both objectives, the objective in itself is not the same as illustrated in eq. (1.4) in the case where the predictive function can be written as an integral which is for example the case for the mean.

$$\underbrace{\int \varphi(y,x)\,\underbrace{p(y \mid x)}_{\text{Objective}}\,\mathrm{d}y}_{\text{By-Product}} \qquad \underbrace{f(x) = \int \varphi(y,x)\,\underbrace{p(y \mid x)}_{\text{By-Product}}\,\mathrm{d}y}_{\text{Objective}}. \tag{1.4}$$

While many of the tools used for prediction are the same as those used in the more traditional approaches, the stark difference in the question being answered warrants different types of theoretical results. As said earlier, we are interested in solving the problem of eq. (1.3), that is, finding the best on average possible f from a family $\mathcal{F}$ of potential functions where our criterion depends on the end goal. Of course, we do not know the distribution of (Y, X) or even the true family of function that encompasses the true f, and therefore cannot directly solve eq. (1.3). We can, however, solve the empirical version of this problem from the data we have at hand, that is

$$\operatorname*{argmin}_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}\left(Y_i, f(X_i)\right), \tag{1.5}$$

which we call the Empirical Risk Minimization (ERM) approach. Given that we do not solve the right problem but only an empirical and restricted version of it, it seems legitimate to ask what guarantees one have that the solution obtained is a good solution. This last question is the foremost problem of *statistical learning theory* and has been approached under many angles, we are, however, taking here the probably approximately correct (PAC) approach: we say that our solution $\hat{f}_n$ of eq. (1.5) is good if it is good according to eq. (1.3) with high probability. That is, given that we only know to compute

$$\mathcal{R}_n(\hat{f}_n) = \min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}\left(Y_i, f(X_i)\right),$$

that the quantity

$$\left|\mathcal{R}_n(\hat{f}_n) - \mathcal{R}(f^*)\right|, \tag{1.6}$$

is small,[25] where $\mathcal{R}(f^*)$ is the minimum of eq. (1.3). Many results of the

24: Respectively by choosing the pinball loss and the cross-entropy loss.

25: As well as $\mathcal{R}_n(\hat{f}_n)$ being small but this is implicit given that it is explicitly constructed that way.

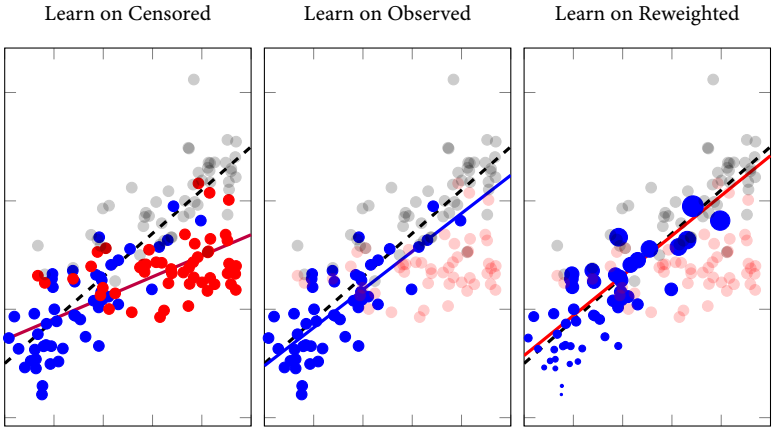


FIGURE 1.8: Learning a linear regression on the raw censored data, the fully observed data only and the reweighted observed data.

form exist, as will be seen in §2.1 but one glaring issue remains that make the ERM approach unsuitable for survival analysis: in eq. (1.5) we do not observe Y_i. It is, however, possible to adapt eq. (1.5) to the survival setting and prove results similar to those already existing in statistical learning theory without censoring.

1.3.1 Censored Prediction

In our setting Y is unobserved and only (T, δ), the censored variable as well as censoring indicator, are observed instead. We show, following the seminal series of Stute (1996, 1993a,b, 1995a,b, 2003); Stute and J.-L. Wang (1993) and work of Dabrowska (1989) that the unobservable quantity of eq. (1.3) can be replaced by the reweighted, but mathematically equivalent

$$\underset{f}{\operatorname{argmin}}\, \mathbb{E}\left[\frac{\delta}{S_C(T- \mid X)}\mathcal{L}\left(T, f(X)\right)\right], \tag{1.7}$$

and corresponding empirical version

$$\underset{f \in \mathcal{F}}{\operatorname{argmin}}\, \frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i}{S_C(T_i \mid X_i)}\mathcal{L}\left(T_i, f(X_i)\right). \tag{1.8}$$

Of course, while the unobservable variable Y has been replaced by the observable quantities T and δ, we now instead involve the unknown survival function S_C. By solving this new reweighted empirical problem instead of relying on the whole censored dataset or only the fully observed individuals, we are able to eliminate the estimator bias which would otherwise result in an underestimation, as can be seen in fig. 1.8. While this change

seems pointless as we exchanged an unknown quantity for another, we know how to estimate S_C as seen in §1.2 and we can instead study

$$\operatorname*{argmin}_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \frac{\delta_i}{\hat{S}_C(T_i \mid X_i)} \mathcal{L}\left(T_i, f(X_i)\right). \tag{1.9}$$

In chapter 2, we show that when using a kernel estimator $\hat{S}_C$ of S_C, we can derive non-asymptotic and nonparametric bounds of the generalization error of eq. (1.6) similar to those of the standard, uncensored, statistical learning literature. As $\hat{S}_C(T_i \mid X_i)$ itself is a random variable involving all the (T_i, X_i) of the training data, here under the form of a sum of independent estimators, the ratio

$$\frac{\delta_i}{S_C(T_i \mid X_i)} \mathcal{L}\left(T_i, f(X_i)\right),$$

is not independent and identically distributed which renders most of the proof techniques involving empirical sums of i.i.d. variables invalid. Instead we rely on the fact that the previous quantity can be written as a ratio of sums in order to linearize it and subsequently treat it as a U-statistic, i.e. a generalization of empirical means, on which concentration results can be applied. This enables us to prove in Theorem 2.9 generalization bounds on the censored ERM problem that are similar to those in the completely observed case:

Theorem (Uniform control of the excess risk)**.** *Suppose that Assumptions 2.1 and 2.4 are fulfilled. There exist constants h_0, M_1, M_2 and M_3 that depend on (A, v), M_Φ, L, K and b only, such that, for all $n \geq 2$ and $\varepsilon \in (0, 1)$, the event*

$$\left|\mathcal{R}(\tilde{f}_n) - \mathcal{R}(f^\star)\right| \leq M_1 \left(\sqrt{\frac{\log\left(M_2/\varepsilon\right)}{n}} + \frac{\left|\log(\varepsilon h^{d/2})\right|}{nh^d} + h^2 \right),$$

occurs with probability greater than $1 - \varepsilon$ provided that $h \leq h_0$, $nh^{2d} \geq M_3 \left|\log(\varepsilon h^d)\right|$.

Moreover, we prove experimentally in §2.5 that the performance obtained on real data by the proposed framework match those expected from the theoretical bounds.

These results, which represent the main contribution of this thesis, have been presented preliminarily at the Machine Learning for Health Workshop at NeurIPS 2018 (Ausset, Portier, and Clémençon [2018]) and are currently under final review for publication at JMLR at the time of writing.

Papers of chapter 2

Guillaume Ausset, François Portier, and Stéphan Clémençon (2018). "Machine Learning for Survival Analysis: Empirical Risk Minimization for Censored Distribution Free Regression with Applications". In: NeurIPS ML4H Workshop. Montreal, Canada

```
@inproceedings{aussetMachineLearningSurvival2018,
  title = {Machine {{Learning}} for {{Survival Analysis}}:
    {{Empirical Risk Minimization}} for
    {{Censored Distribution Free Regression}} with {{Applications}}},
  author = {Ausset, Guillaume and Portier, François and Clémençon, Stéphan},
  date = {2018},
  location = {{Montreal, Canada}},
  url = {https://hal.archives-ouvertes.fr/hal-02287991},
  eventtitle = {{{NeurIPS ML4H Workshop}}}
}
```

Guillaume Ausset, Stéphan Clémençon, and François Portier (2021a). "Empirical Risk Minimization under Random Censorship". *under revision in Journal of Machine Learning Research*. arXiv: 1906.01908

```
@article{aussetEmpiricalRiskMinimization2021,
  title = {Empirical {{Risk Minimization}} under {{Random Censoring}}},
  shorttitle = {Empirical {{Risk Minimization}} under {{Random Censoring}}},
  author = {Ausset, Guillaume and Clémençon, Stéphan and Portier, François},
  date = {2021},
  journaltitle = {Journal of Machine Learning Research (Provisional)},
  url = {http://arxiv.org/abs/1906.01908}
}
```

1.3.2 Flexible Estimators of the Survival

While the results presented in chapter 2 give strong theoretical justifications for the use of IPCW ERM, the performance still depends heavily on the quality of the weights $\delta_i/S_C(T_i|X_i)$ and therefore of the estimator of S_C. Beyond the use in IPCW regression, estimators of the survival are of interest to the survival community by themselves and many flexible such estimators have been proposed through the years. In chapter 3, based on Ausset, Ciffreo, et al. (2021), we study a particular type of estimator of S constructed from a generative model of the variable of interest, i.e. Y in the general survival setting or C for the IPCW weights, with a tractable likelihood. By modelling Y as the transformed variable

$$Y = m_\theta(Z, X), \tag{1.10}$$

where m_θ is a highly flexible family of neural networks parametrized by $\theta \in \Theta$ and Z is a simple known distribution. We are able to find the

optimal choice of m_θ by maximizing the censored log-likelihood

$$\sum_{i=1}^{n}\Big(\delta_i p_{Y,\theta}\left(T_i \mid X_i\right) + (1-\delta_i)\, S_{Y,\theta}\left(T_i \mid X_i\right)\Big), \tag{1.11}$$

where $p_{Y,\theta}$ and $S_{Y,\theta}$ are the density and survival function of Y as parametrized by eq. (1.10). The parametrization given by eq. (1.10) is a type of generative model first introduced under the name *normalizing flow* by Rezende and Mohamed (2015). Its usefulness resides in the fact that $p_{Y,\theta}$ can be derived from p_Z by means of the change of variable formula

$$\log p_{Y,\theta}(t \mid X) = \log p_Z(z) - \log\left|\det \frac{\partial m_\theta}{\partial z}\right|.$$

Similarly, we show in chapter 3 that it is also possible to retrieve the survival $S_{Y,\theta}$ by adopting the continuous formulation of eq. (1.12) (see R. T. Q. Chen et al. [2018]),

$$\begin{aligned}\frac{\partial}{\partial t}\begin{bmatrix}\mathbf{z}_\theta(t,X)\\ \log p(y \mid X) - \log p(\mathbf{z}_\theta(t,X))\end{bmatrix} &= \begin{bmatrix} m_\theta(\mathbf{z}_\theta(t,X),t,X)\\ -\operatorname{tr}\dfrac{\partial m_\theta}{\partial \mathbf{z}}\end{bmatrix},\\ \begin{bmatrix}\mathbf{z}_\theta(1,X)\\ \log p(y \mid X) - \log p(\mathbf{z}_\theta(1,X))\end{bmatrix} &= \begin{bmatrix} y\\ 0\end{bmatrix},\end{aligned} \tag{1.12}$$

making it possible to compute as well as differentiate (see Rackauckas, Ma, Dixit, et al. [2018], for differentiability of solutions of ordinary differential equations (ODEs)) all the quantities present in eq. (1.11).

Despite the high computational cost of the method proposed we show that compared to existing neural approaches such as DeepCox (Nagpal et al. [2021]) or DeepHit (C. Lee, Zame, et al. [2018]; C. Lee, Yoon, and Schaar [2020]), this continuous normalizing flow (CNF) approach performs competitively on classical regression tasks but also enables new applications. As a generative model, the CNF approach gives the ability to efficiently sample conditional observations, a very useful characteristic in finance where stress-tests and simulations are regulatory requirements; but also for applications where complex dependencies have to be modelled and simulated by means of Monte Carlo. This last application, given its particular relevance to finance is studied in greater details in chapter 5.

While the advantages of very flexible, and very expensive, neural network based generative models of the survival are undeniable, the computational cost can be hard to justify when considering the relatively high performance of simpler, and nearly free in comparison, methods such as random survival forest (RSF) (Ishwaran and Kogalur [2007]) or even Cox models (D. R. Cox and Oakes [1984]); even if one could argue that most of the cost is incurred during training and amortized during inference.

Despite the fact that the CNF method is inherently more expensive, it is still possible to mitigate the computational burden by reducing the size of the neural network, this bandaid can work only if the number of dimensions of X itself is lowered at the same time. Therefore, in order for the method proposed to be of practical interest, a robust way of reducing the dimension of X is needed.

PAPERS OF CHAPTER 3

Guillaume Ausset, Tom Ciffreo, et al. (2021). "Individual Survival Curves with Conditional Normalizing Flows". In: *DSAA'21*. IEEE International Conference on Data Science and Advanced Analytics

```
@inproceedings{aussetIndividualSurvivalCurves2021,
  title = {Individual {{Survival Curves}}
    with {{Conditional Normalizing Flows}}},
  booktitle = {{{DSAA}}'21},
  author = {Ausset, Guillaume and Ciffreo, Tom
    and Clémençon, Stéphan and Portier, François and Papin, Timothée},
  date = {2021},
  eventtitle = {{{IEEE International Conference}} on {{Data Science}}
    and {{Advanced Analytics}}}
}
```

1.3.3 Dealing with High Dimension

The final major contribution of this thesis presented in chapter 4 is a response to the previously mentioned need for a robust dimension reduction technique. The goal of dimensionality reduction is to find a lower-dimensional space which captures most of the information present in the original space. The very notion of *information* is left here intentionally fairly vague as, depending on the task or the specific needs of the problem of interest, it can change drastically thus leading to very different notions of reduction of dimension.

We can very roughly divide the types of tasks of interest in two distinct[26] groups: *unsupervised* and *supervised*. By *unsupervised* we mean tasks where the object of interest is X itself which is considered to be the only quantity observed.[27] Without any auxiliary information, the most natural way of framing the problem is therefore simply to view it as a reconstruction problem that is finding a lower-dimensional space obeying some additional constraints such that this new space minimizes some notion of distance to the original space. This is for example the approach taken by principal components analysis (PCA) where the square loss of a projection onto a subspace of dimension l is minimized.[28] Similarly, variational autoencoders (VAEs) find the latent representation Z of lower

26: But with significant overlap.

27: Even then, some people see this task as a *self-supervised* problem, that is where the covariates are X and the dependant variable also X.

28: Incidentally this is equivalent to maximizing the variance on the projected subspace.

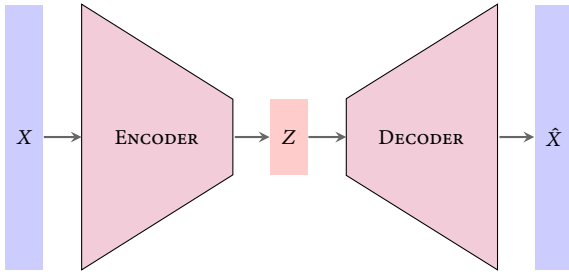


Figure 1.9: Simplified VAE.

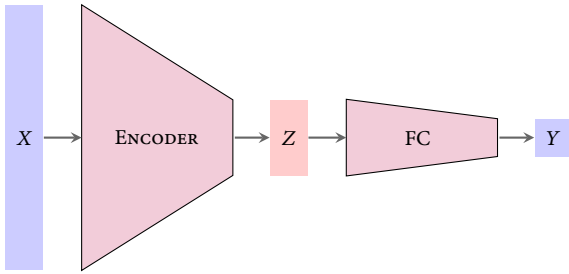


Figure 1.10: Last layer of DNN as representation.

dimension l by jointly learning the embedding function or encoder enc and decoding function dec that minimizes the reconstruction error

$$\|X - \text{dec} \circ \text{enc}(X)\|,$$

which can be seen as a generalization of PCA, when taking enc $= P$ a projection and dec $=$ id[29] as illustrated in fig. 1.9.[30] On the other hand, *supervised* dimensionality reduction techniques are concerned with finding good representations of X when (X, Y) is observed and the prediction of Y given X is the task of interest. It is for example possible to see linear discriminant analysis (LDA) as a supervised extension of PCA[31] which instead of finding the projection that maximizes the variance, finds the projection that maximizes the class separation. Similarly, by analogy with the VAE approach, instead of finding a representation suitable for reconstruction, it is usual in the supervised setting to find a representation suitable for prediction, or classification, by simply taking the last layer of a deep neural network (DNN), before the fully connected (FC) output layer, as our lower dimensional embedding of interest as represented in fig. 1.10.

Another approach to the supervised problem is to consider as important the variables that have an impact on the output

$$Y = f(X) + \varepsilon,$$

which fairly naively can be thought as finding the direction of non-zero derivatives. If we limit ourselves solely to variable selection, that is restricting the possible directions only to the axes, then the problem is simply of

29: This is not an exhaustive review, more details are given in chapter 4. Note that while all these techniques can be seen from a pure reconstruction point of view, i.e. pure optimization problems, they also admit probabilistic interpretations.

30: The real formulation of VAEs is probabilistic and variational in essence as the formulation given here will overfit the data.

31: LDA is often seen as a classification algorithm but it can be used for dimension reduction.

finding the non-zero elements of the gradient. This approach has been studied and justified in the literature under the *single* and *multi index* framework where

$$Y = f(TX) + \varepsilon,$$

with T a projection matrix. Under this model it is clear that the effective dimension reduction (EDR) subspace defined by T is spanned by the gradient ∇r of $r(x) = f(Tx)$. Several approaches to this specific problem have already been proposed in the literature and are described in detail in §4.1 but non combine non-asymptotic and uniform bounds on the error of both the gradient and regression function itself with a k-NN approach when the gradient is supposed sparse. The k-NN approach is particularly attractive in practice as it is not only easy to calibrate and understand for laymen, but also prevents any pathological cases where too few examples are present in a chosen neighbourhood. We give in chapter 4 a local linear least absolute shrinkage and selection operator (LASSO) formulation of the problem of the form

$$\underset{(r,\beta)\in\mathbb{R}^{d+1}}{\operatorname{argmin}} \sum_{i\in\widehat{i}_k(x)} \left(Y_i - r - \beta^\intercal (X_i - x)\right)^2 + \lambda\|\beta\|_1, \tag{1.13}$$

and show in Theorem 4.1 that it is possible to exploit the supposed sparsity of the gradient to improve the bounds on the error, with similar results on the regression function itself in Theorem 4.2.

Theorem. *Suppose that Assumptions 4.1 to 4.4 are fulfilled. Let $n \geq 1$ and $k \geq 1$ such that $C_k \leq C_0$ and take*

$$\lambda = C_k \left(\sqrt{2\sigma^2 \frac{\log(16d/\delta)}{k}} + LC_k^2 \right).$$

Then, we have with probability larger than $1-\delta$,

$$\left\|\tilde{\nabla}_k r(x) - \nabla r(x)\right\|_2 \leq 24^2 \sqrt{|S_x|} \left(C_k^{-1} \sqrt{\frac{2\sigma^2 \log(16d/\delta)}{k}} + LC_k \right),$$

as soon as

$$C_1 \, |S_x| \log\left(\frac{dn}{\delta}\right) \leq k \leq C_2 n,$$

where $\tilde{\nabla}_k r(x)$ is the second component of the solution of eq. (1.13), C_0, C_1, C_2 and L are universal constants, C_k is a constant defined in detail in Theorem 4.1 and $|S_x|$ is the number of non-zero components of $\nabla r(x)$.

Most of the examples of dimension reduction given earlier, such as PCA or LDA, assume that the important variables are the same for all individuals and are therefore treated as a preprocessing step applied to the whole

dataset before any further analysis. There is, however, no reason for such a strong hypothesis to be true: if the dataset includes for example men and women, it seems dubious at best to think that the same variables are important for both sexes and an adaptive variable selection method would be expected to perform better. Similarly, when comparing the characteristics of clients of a loan, it seems desirable to take into account different characteristics if the client is a multinational or a small local company. As our method is local, that is the gradient is estimated at a specific x and therefore retrieves the variables of importance in a neighbourhood of x, it is possible to select different relevant variables in different regions of the space. In §4.5, not only do we show how to find *globally* important variables by aggregating the gradients of all observations but we also propose a tree-based method exploiting the local gradient information in order to improve performance.

Finally, while the theoretical analysis in the chapter is done while momentarily ignoring all forms of censoring for reasons of simplicity, it is still an ERM problem on which we can apply the IPCW approach of chapter 2, for example to identify the genes responsible for the survival of cancer as done in §4.5.2.

Papers of chapter 4

Guillaume Ausset, Stéphan Clémençon, and François Portier (2021b). "Nearest Neighbour Based Estimates of Gradients: Sharp Nonasymptotic Bounds and Applications". In: *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics*. Ed. by Arindam Banerjee and Kenji Fukumizu. Vol. 130. Proceedings of Machine Learning Research. PMLR, pp. 532–540

```
@inproceedings{aussetNearestNeighbourBased2021,
  title = {Nearest Neighbour Based Estimates of Gradients:
    {{Sharp}} Nonasymptotic Bounds and Applications},
  booktitle = {Proceedings of the 24th International Conference
    on Artificial Intelligence and Statistics},
  author = {Ausset, Guillaume and Clémencon, Stephan and Portier, François},
  date = {2021},
  series = {Proceedings of Machine Learning Research},
  volume = {130},
  pages = {532--540},
  publisher = {{PMLR}},
  url = {http://proceedings.mlr.press/v130/ausset21a.html}
}
```

1.4 Outline of this Manuscript

This dissertation is organized as follows:

- Chapter 2 deals with the empirical risk minimization framework in the presence of censoring. Non-asymptotic and uniform upper bounds of the generalization error are proven, while the end of the chapter is dedicated to numerical experiments to justify the validity of the approach beyond the theoretical results.
- Chapter 3 introduces normalizing flows for survival analysis, a generative model with a tractable likelihood. Several applications serve as an example of the utility of such an approach in the classical survival setting while the motivation for the generative formulation is studied in greater details in chapter 5.
- Chapter 4 treats of the problem of high dimension through the scope of variable selection. An estimator of the gradient is introduced and non-asymptotic bounds on the error of both the supposedly sparse gradient and the regressor are derived. As the gradient is itself useful beyond simple variable selection, several examples of zeroth-order optimization as well as disentanglement are given.
- Chapter 5 treats of survival analysis in finance through the specific case of securitization, one of the core business of BNP Paribas CIB. We motivate the generative model introduced in chapter 3 through the use of hierarchical multilevel modelling.

Beyond the main text which introduces the contributions of this thesis, some additional details are given in the appendix.

- Classical proofs not strictly needed for the comprehension of the main results are given in appendix A for reference.
- An historical overview of survival analysis is given in appendix B as a distraction and intermission to the technical proofs.
- Appendix C contains a French translation of this introduction.

Prediction and Censoring | 2

2.1 Introduction

We have quickly introduced the survival analysis setting in the previous chapter and given a quick overview of the mathematical objects involved.[33] We have however, until now, mostly concerned ourselves with constructing estimators of the density, survival or hazard, and then calibrating[34] them in order to obtain the parameters corresponding to the best possible representation of the hypothesized distribution of the observed censored data. We will refer,[35] to the approach of first making a hypothesis on the distribution of the data and then finding the best parameters in order to have our hypothesis and the observed data be in agreement with the goal of *understanding* the process that generated the data, as the *statistical approach*. We have started by describing the *statistical approach* in the previous chapter as it is not only historically the first approach developed but more importantly because it is, more often than not, the approach that corresponds most closely with the way scientists work, and is capable of answering the questions those same scientists have. While we frame mathematically our problem as predicting the time of death of a patient, this is often not of any direct interest to the researcher. A medical researcher or a biostatistician is probably more interested in *understanding* the causes and mechanisms that lead to the death, than the death itself, and is therefore more interested in the inferred parameters of their hypothetical law than the mean, mode or any other output of said law. A medical researcher working on the survival of trauma patients suffering from major blood loss may for example decide to model the survival of patients given their characteristics X_i using a proportional hazard model i.e. such that

$$\lambda(t \mid X_i) = \lambda_0 \exp(\beta_1 X_{i1} + \ldots + \beta_p X_{ip}) = \lambda_0 \exp(\beta^\intercal X_i).$$

This model, commonly known as the Cox proportional hazard model after his inventor Sir David Cox (D. R. Cox [1972]), entirely define the law of the time of death and can therefore be used to estimate the mean time of death or any other statistic of interest. The object of interest for our medical researcher however is here β, the parameters, which can then be used to

33: See §1.2 if you have missed it.

34: Or *fitting* them, if more familiar with the machine learning terminology

35: Hopefully without offending anybody, as there are as many definitions of statistics and machine learning as there are statisticians and machine learners.

derive valuable insights such as the relative importance of the different biomarkers, or to compare the relative survival of two populations.[36] When considering this point of view, where the true object of importance is the parameters and not the model itself, it is not surprising that most of the survival analysis literature, mostly driven by medicine, has historically focused on studying the properties of the estimated parameters. From this rich literature of the statistical aspects of survival analysis have emanated a vast and comprehensive collection of results quantifying the asymptotic convergence of the various estimators proposed. We refer the readers interested in this approach to Klein and Moeschberger (2003) while in this thesis in general and this chapter in particular, we adopt a wildly different point of view of the survival analysis setting which leads us to wildly different results.

36: If the question of interest is "Is the survival of a different than b?" then there are more proper statistical tests.

In many practical settings, completely antithetically to the researchers of the previous paragraph, we are not interested in understanding a phenomenon at all and instead only interested in our ability to predict it. While a physicist may be interested in understanding the *why* and *how* of gravity, a plane designer may only be interested in making sure it flies without a care in the world for gravity.[37] This shift in objectives coincides with the shift of practitioners in recent years: while analysing data was once seen as a tool for scientists, *data analysis* is born from the necessity for laymen to exploit the ever-growing mass of data the industry has been accumulating. As the questions themselves are different, new tools and results are necessary in order to support those new needs. We will focus here on one of those questions: regression, which we will refer to as the *prediction* problem even though it, in reality, only represents a fraction of the field.[38]

37: It may be apparent that I know nothing about aeronautics, plane designers quite certainly care about gravity. I hope.

Regression answers the question: "*given an input describing an individual, what is the output?*". In order to answer this question, we need to formalize the question. We will, following the earlier notations, define our output as Y, representing for example the time until a certain event, and input as X which represents the characteristics of the individual in question. Our question is then how to map X to Y, or more formally, what is the function f that maps X to Y. As there are is reason for a mapping

$$Y = f(X),$$

to exist or be unique; not only because most processes cannot be exactly described given the limited input we have access to but because even in the cases where it could in theory be possible (for a physical phenomenon for example) we usually assume some level of uncertainty due to errors of measurement.[39] Instead, we try to find the best possible, or at least good enough, mapping to answer the question. *Best possible* is a very subjective term and is usually best defined specifically for the task at hand by the

38: "Prediction" depends on what exactly one wants to predict. Classification, structured prediction and many other problems are also part of the larger framework we describe here through the scope of regression.

39: We usually assume errors in the measurements of the output such that $Y = f(X) + \varepsilon$, but errors can also occur in the input. Mathematically this doesn't change our model.

practitioner as it depends on the business or industrial imperatives but in the most general setting, with some undetermined notion of fitness $\mathcal{L}$, or *loss*[40] in the machine learning literature we can rewrite our question as

$$f^\star = \operatorname*{argmin}_f \mathbb{E}\left[\mathcal{L}(Y, f(X))\right], \tag{2.1}$$

40: Mathematicians traditionally prefer minimizing quantities to maximizing them.

where we have taken the expectation in order to account for the uncertainty in both Y and X. We will refer to the previous quantity that is being minimized as the population risk $\mathcal{R}(f^\star)$, that is:

$$\mathcal{R}(f) = \mathbb{E}\left[\mathcal{L}(Y, f(X))\right], \tag{2.2}$$

where the expectation is taken under the true, unknown, distribution. Of course there is no hope recovering the true minimizer of the population risk as we cannot feasibly minimize over all possible functions, we instead have to constrain ourselves to some class $\mathcal{F}$ of candidates $f \in \mathcal{F}$ where $\mathcal{F}$ can be the class of linear functions, the class of polynomials or any other class of function suited to the problem. Contrary to a large part of the statistical literature, we will not assume that the class of function $\mathcal{F}$ contains the *true* mapping $f^\star$. The previous problems can then be written in its final form as

$$\bar{f} = \operatorname*{argmin}_{f \in \mathcal{F}} \mathbb{E}\left[\mathcal{L}(Y, f(X))\right], \tag{2.3}$$

which can be rewritten in terms of population risk:

$$\bar{f} = \operatorname*{argmin}_{f \in \mathcal{F}} \mathcal{R}(f),$$

where this last quantity should be read as *"find the best possible function $f^\star$ amongst all the allowed functions $\mathcal{F}$ that minimizes the loss*[41] *$\mathcal{L}$ in average"*.

In practice, we cannot write the population risk given earlier, but we instead have access to realizations of the law of (Y, X). In order to simplify things, we make the not so constraining hypothesis that those realizations are all identically distributed and independent. We denote by,

41: Our notion of un-fitness

$$\mathcal{D}_n \overset{\text{def}}{=} \{(Y_1, X_1), \ldots, (Y_n, X_n)\},$$

the training[42] set comprising of the n independent and identically distributed copies of (Y, X) available to us. Given the hypothesis and data available to us, instead of solving the problem eq. (2.3) we solve its empirical variant as it is the only one available to us:

42: Practitioners will often split their data in *training*, *validation* and *testing* sets.

$$\hat{f}_n = \operatorname*{argmin}_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}(Y_i, f(X_i)). \tag{2.4}$$

If we define the empirical risk as

$$\mathcal{R}_n(f) = \sum_{i=1}^{n} \mathcal{L}(Y_i, f(X_i)), \tag{2.5}$$

we can then state the previous problem in terms of minimizing the empirical risk instead of the population risk:

$$\hat{f}_n = \underset{f \in \mathcal{F}}{\operatorname{argmin}}\, \mathcal{R}_n(f).$$

The prediction problem can then be solved by following a simple schema:

1. Select a loss $\mathcal{L}$ appropriate to the task at hand,
2. Select a candidate family $\mathcal{F}$,
3. Select a minimization strategy to solve eq. (2.5),
4. Predict new Y_k for unobserved X_k as $Y_k = \hat{f}_n(X_k)$.

Each one of the previous points lead to its own branch of research and is deserving of a thesis in its own right but we will here simply focus on justifying the validity of such an approach.

Remember that the real problem of interest is defined in eq. (2.2) but not only has been constrained to a limited choice of candidates $\mathcal{F}$ in eq. (2.3) but more importantly is solved in practice on an empirical, and therefore random, estimator of this same risk in eq. (2.5). It is therefore legitimate to ask the question: "*How far is the estimated function $\hat{f}_n$ from the true function $f^\star$?*". An attentive observer may notice that this question is actually more general than what we truly are interested in: our objective is expressed in terms of loss and we are not interested in how close $\hat{f}_n$ and $f^\star$ are,[43] but only in $\mathcal{R}(\hat{f}_n)$. Ideally, as we know that the best possible risk we can achieve is $\mathcal{R}(f^\star)$, we would want to make sure $\mathcal{R}(\hat{f}_n) - \mathcal{R}(f^\star)$ is small. This is, however, an unrealistic goal as we can at best only hope to measure and control the risk restricted on the candidate set $\mathcal{F}$. The risk eq. (2.3) represents the best possible risk we, the practitioners, can hope to achieve short of being omniscient. We therefore will only try to control the excess risk, defined as

43: While it would indeed give us the answer, this is a much harder problem.

$$\begin{aligned}\mathcal{E}(\hat{f}_n, \mathcal{F}) &= \mathcal{R}(\hat{f}_n) - \inf_{\mathcal{F}} \mathcal{R}(f) \\ &= \mathcal{R}(\hat{f}_n) - \mathcal{R}(\bar{f}).\end{aligned} \tag{2.6}$$

Such guarantees exist and are the basis of the justification of the field of machine learning; but before introducing them, we bring several facts to the attention of the reader. First, eq. (2.4) depends on the training set $\mathcal{D}_n$ and we therefore expect our guarantees on the excess risk to depend at the very least on its size $|\mathcal{D}_n| = n$. Secondly, the same problem depends on the class of candidates $\mathcal{F}$ and naively, but quite naturally, one would

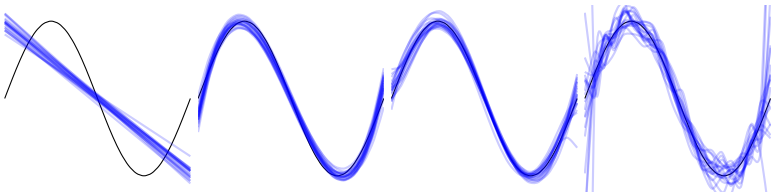


FIGURE 2.1: Learning $\sin(x)$ over $[0, 2\pi]$ with polynomials of degrees 1,3,5 and 20.

expect the quality of the solution to increase with the *complexity* or *richness* of $\mathcal{F}$. However some quick experiments quickly prove this idea wrong. Figure 2.1 represents for example the result of learning a simple function with an increasingly complex family of candidates: the quality of the results (in terms of prediction i.e. the loss achieved on new unseen examples) increases with the complexity of the family $\mathcal{F}$ until a certain point until it deteriorates, potentially unboundedly so.

A simple way to think of the previous phenomenon, where highly flexible families of candidate functions are able to increasingly well fit the training set while at the same time being inadequate for prediction on new observations, is to think in terms of "bias-variance[44]" decomposition of the risk. After simple algebraic manipulations, we can obtain the following decomposition of the excess risk of eq. (2.6):

$$\mathcal{E}\left(\hat{f}_n\right) = \underbrace{\left(\mathcal{R}(\hat{f}_n) - \mathcal{R}_n(\hat{f}_n)\right)}_{A} + \underbrace{\left(\mathcal{R}_n(\hat{f}_n) - \mathcal{R}_n(\bar{f})\right)}_{B \leq 0} + \underbrace{\left(\mathcal{R}_n(\bar{f}) - \mathcal{R}(\bar{f})\right)}_{C},$$

where B is negative by construction. The term C can be dealt straightforwardly by means of concentration inequalities (see appendix A) by noticing that it can be written as

$$C = \left(\frac{1}{n}\sum_{i=1}^{n} \mathcal{L}\left(Y_i, \bar{f}(X_i)\right)\right) - \mathbb{E}\left[\mathcal{L}\left(Y_i, \bar{f}(X_i)\right)\right].$$

The term of interest, or at least the one requiring us to develop additional techniques and which motivates the field of statistical learning in general and this chapter in particular is therefore the term A:

$$A = \mathbb{E}\left[\mathcal{L}(Y_i, \hat{f}_n(X_i))\right] - \left(\frac{1}{n}\sum_{i=1}^{n} \mathcal{L}(Y_i, \hat{f}_n(X_i))\right). \tag{2.7}$$

The reason this term is fundamentally different to the previous term, despite looking so similar, is that now $\hat{f}_n$ is itself a random quantity. In order to be dealt with, we need concentration bounds similar to those reminded in appendix A, that is tail bounds uniform in f.

Coming back to the toy example of fig. 2.1, we can visualize in fig. 2.2 this tradeoff by plotting the error rate on new unseen observations with

44: A true bias-variance interpretation can be obtained if the loss is the squared error loss. Here it is only an analogy.

45: It may not be apparent on the plot due to plotting innacuracies but the error goes to 0.

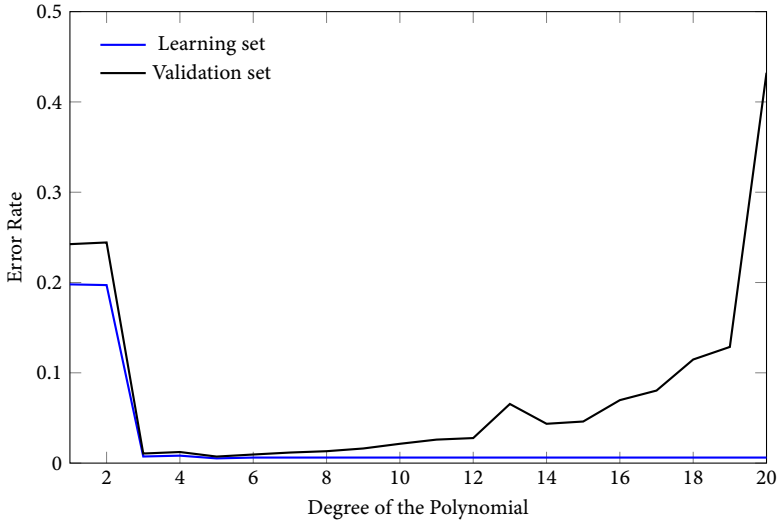


FIGURE 2.2: Overfitting on the problem fig. 2.1.

respect to the degree of the polynomial. While the error on the training set decreases consistently,[45] the error on new previously unseen observations starts increasing after a certain threshold. In order to quantify the previous phenomenon, we need a way to objectively measure the complexity of a family of functions. For a parametric family[46] it can be tempting to use the number of parameters as a measure of the complexity of the class of functions. While one could argue it can be a decent proxy,[47] it is clear that it is close to useless to compare different families: there is no reason for a linear regression with 1000 parameters to be more powerful or more expressive than a support vector machine with 100 parameters. At one extreme, we could take as candidate family the space of all functions, which even without a formal definition of complexity we will accept as *highly complex*, and achieve a perfect error on the training set (just take the function that maps X_i to Y_i for all $i \in [\![1, n]\!]$ and does *something else* for the rest of the space.) but arbitrarily bad on new examples. On the other hand, by taking the class of constant functions, it is clear that the error on the training set will be lackluster but the performance on new examples will be strictly equivalent in average. There is therefore a need for a robust definition of complexity on which to prove results pertaining to generalization.

46: Parametric families represent the bulk of the families considered as those are easy to reason about and are amenable to optimization in practice.

47: I disagree that it is valid proxy. While still the main criterion in fields like deep-learning, it is clear that adding useless parameters do not change the complexity in any way.

2.1.1 Functional Complexity

As quickly evoked earlier, the quantity of interest to control the excess risk resemble greatly the quantities usually encountered when deriving

the laws of large numbers and tail bounds. We, however, need to derive a *uniform* law of large numbers, which we will be able to do by first defining a notion of complexity of functions as teased earlier.

For a fixed sample of observations $\mathcal{D}_n^x = \{x_1, \dots, x_n\}$, we define the set

$$\mathcal{F}(\mathcal{D}_n^x) = \{(f(x_1), \dots, f(x_n)) \mid f \in \mathcal{F}\},$$

of points in $\mathbb{R}^n(\mathcal{D}_n^x)$ that can be achieved as mappings of functions of $\mathcal{F}$. We can then define the empirical Rademacher complexity of $\mathcal{F}$ as

$$\mathfrak{R}\left(\mathcal{F}(\mathcal{D}_n^x)\right) \stackrel{\text{def}}{=} \mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i f(x_i)\right|\right], \tag{2.8}$$

where the (ε_i) are independent and identically distributed Rademacher variables i.e. variables such that $\mathbb{P}(\epsilon = 1) = 1/2$ and $\mathbb{P}(\epsilon = -1) = 1/2$. As the previous quantity is a random variable, we define the Rademacher complexity of $\mathcal{F}$ as

$$\mathfrak{R}_n(\mathcal{F}) \stackrel{\text{def}}{=} \mathbb{E}\left[\mathfrak{R}\left(\mathcal{F}(\mathcal{D}_n^x)\right)\right] = \mathbb{E}\left[\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i f(x_i)\right|\right]\right]. \tag{2.9}$$

The Rademacher complexity $\mathfrak{R}_n(\mathcal{F})$ represents the average maximum correlation between a dataset class of function and noise. This interpretation motivates the Rademacher complexity as a useful definition of functional complexity: it measures the ability of a class of function to model random noise. Going back to our previous polynomial example, we can see how this class of function behaviour matches our definition of complexity: a polynomial of sufficiently high degree is able to capture any "pattern" in the data, even noise, and is therefore entirely useless to make any predictions.[48]

In order to control the variations of the process of eq. (2.7) we need a uniform version of the classical Glivenko-Cantelli theorem:

48: We can see how our problem matches the counterfactual problem encountered in science: any model that is sufficiently powerful to explain *everything*, and can therefore never be *falsified* is useless.

Definition 2.1 (Uniform Glivenko-Cantelli)**.** We say that $\mathcal{F}$ is a Glivenko-Cantelli class of function if

$$\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} \stackrel{\text{def}}{=} \sup_{\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n} f(X_i) - \mathbb{E}\left[f(X)\right]\right| \xrightarrow[n\to\infty]{\mathbb{P}} 0,$$

where $\mathbb{P}_n = \sum_{i=1}^{n}\delta(X_i)$ is the empirical distribution of the data.

It is then possible to state a uniform concentration bound in terms of Rademacher complexity.

Proposition 2.1 (Uniform Glivenko-Cantelli - Rademacher, see e.g. Wainwright (2019))**.** *For any class $\mathcal{F}$ such that $\|f\|_\infty \le b$ for all $f \in \mathcal{F}$ and $n \ge 1$ we have with probability greater than $1 - \varepsilon$*

$$\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} \le 2\mathfrak{R}_n(\mathcal{F}) + \sqrt{\frac{2b^2\log(1/\varepsilon)}{n}}.$$

The proof uses common techniques, some of which we will use in the proofs of the results of the next section. Generally, most of the proofs proceed according to the following scheme:

1. Symmetrization using a ghost sample.
2. Symmetrization with a Rademacher process.
3. Conditioning on $\mathcal{D}_n$.
4. Use of a concentration inequality.

The previous result[49] is enough to obtain the guarantees that are typically expected in machine learning: for a fixed sample size n and some estimator $\hat{f}_n$ that we do not assume to be in the same class as the *true* function, we are able to control how far the risk we measured on the training set is to the true population risk we will achieve on *average* on new examples.

49: This type result is usually referred to as *probably approximatively correct*: for a certain level of approximation we can only guarantee a certain level of *certainty* in the result.

2.1.2 Vapnik-Chervonenkis Dimension

We will quickly introduce here the useful notion of complexity of class of functions known as the VC dimension or VC complexity after its inventors V. Vapnik and A. Chervonenkis[50] (Vapnik [1998, 2000]), which we will heavily employ in the subsequent proofs of both chapters 2 and 4. While chronologically anterior, it can be seen (when skipping the usual middle step of polynomial complexity[51]) as an extension of the Rademecher complexity. We have shown in the previous subsection how the Rademacher complexity can be used to derive uniform tail bounds to control the excess risk, but deriving the Rademacher complexity is tedious; we therefore introduce another less powerful[52] definition of complexity that happens to be an upper-bound of the Rademacher complexity. Moreover, unlike the Rademacher complexity, the VC dimension of a class of functions can often easily be proven to exist if the class of function can be built from other simple VC classes using common operations which we will clarify later.

50: Researchers in the field of machine learning like to joke that everything in the field has already been discovered by Soviet mathematicians. This further strengthen this hypothesis.

51: Unfortunately, the more convenient definitions of complexity results in bounds that are not as tight. Polynomial complexity would sit between Rademacher and VC.

52: In term of the tightness of the bounds.

Definition 2.2 (Vapnik-Chervonenkis' dimension)**.** We say that a family of indicator functions *shatter* $\mathcal{D}_n$ if all the dichotomies of $\mathcal{D}_n$ are attainable by that family. That is, in the case of binary classification, that all 2-partitions of $\mathcal{D}_n$ are possible, that is

$$|\mathcal{F}(\mathcal{D}_n^x)| = 2^n.$$

That is, we say that a family of indicator functions is of VC dimension $\nu(\mathcal{F})$ if there exists a set of point $\mathcal{D}_{\nu(\mathcal{F})}$ of size $\nu(\mathcal{F})$ that can be shattered but there doesn't exist a set of point $\mathcal{D}_{\nu(\mathcal{F})+1}$ of size $\nu(\mathcal{F}) + 1$ that can also be shattered by that same family. The VC dimension is therefore the highest number of dichotomies that a family of indicators can achieve. For a family $\mathcal{F}$ of functions f_ω indexed by ω, the VC dimension of $\mathcal{F}$ is defined as the VC dimension of the family $\left(\mathbb{1}_{f_\omega(\cdot)-\beta>0}\right)_{\omega,\beta}$ indexed by ω and β.

The interest of the VC dimension comes from the following theorem that we will not prove:

Proposition 2.2 (VC bound of the Rademacher complexity)**.** *For any class $\mathcal{F}$ and training sample $\mathcal{D}_n$, we have*

$$\mathfrak{R}_n(\mathcal{F}) \leq \sqrt{\frac{2v(\mathcal{F})\log(n)}{n}}.$$

In particular, we can restate Proposition 2.1 in terms of VC dimension:

Corollary 2.3 (Uniform Glivenko-Cantelli - VC dimension)**.** *For any class $\mathcal{F}$ such that $\|f\|_\infty \leq b$ for all $f \in \mathcal{F}$ and $n \geq 1$ we have with probability greater than $1-\varepsilon$*

$$\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} \leq 2\sqrt{\frac{2v(\mathcal{F})\log(n)}{n}} + \sqrt{\frac{2b^2\log(1/\varepsilon)}{n}}.$$

From Corollary 2.3 we can clearly retrieve the expected rate of convergence for this type of problem. As mentioned earlier, one key advantage of the VC dimension is that it is closed under a number of basic operations that we summarize in the two following propositions:

Proposition 2.4 (VC preservation)**.** *If $\mathcal{F}$ and $\mathcal{S}$ are set classes of finite VC dimensions $v(\mathcal{F})$ and $v(\mathcal{S})$ then the following classes of functions also have finite VC dimensions:*

1. *The set class $\mathcal{F}^c \overset{\text{def}}{=} \{F^c \mid F \in \mathcal{F}\}$.*
2. *The set class $\mathcal{F} \cup \mathcal{S} \overset{\text{def}}{=} \{F \cup S \mid F \in \mathcal{F}, S \in \mathcal{S}\}$.*
3. *The set class $\mathcal{F} \cap \mathcal{S} \overset{\text{def}}{=} \{F \cap S \mid F \in \mathcal{F}, S \in \mathcal{S}\}$.*

Proposition 2.5 (Subgraph VC dimension)**.** *Let $\mathcal{F}$ be a vector class of functions $f : \mathbb{R}^d \mapsto \mathbb{R}$ with $\dim(\mathcal{F}) \leq \infty$. Then the subgraph class of $\mathcal{F}$ has VC dimension of at most $\dim(\mathcal{F})$.*

In practice, we will not make use of the previous historical formulation of the VC dimension in terms of shattering number but will instead prefer to express it in terms of packing and covering number (see Wainwright [2019], chapter 5) as it relates directly to the notion of metric entropy (see Kolmogorov [1955, 1956]; Tikhomirov [1957], and others in the *russian school*[53], as well as Definition 2.3.) and enables more natural proofs.

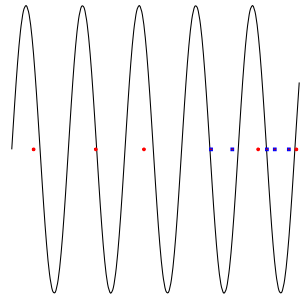


FIGURE 2.3: $\sin(\omega x)$ admits a single parameter but has infinite VC dimensions.

53: Советская формулировка или альтернативная советская формулировка. Множество вариантов.

2.1.3 Risk Minimization and Statistical Learning

We have quickly described in this section what are the results that are expected in machine learning, i.e. bounds controlling the excess risk and how they are usually dealt with in the literature. A large class of interesting problems can be written in terms of minimizing a problem such as in eq. (2.3) and the practical solution then boils down to replacing the unknown distribution $\mathbb{P}$ in the population risk functional $\mathcal{R}(\cdot)$ with the empirical distribution $\mathbb{P}_n$ of the (X_i, Y_i)'s. The class $\mathcal{F}$ of predictive functions is supposed to be of controlled complexity (e.g. of finite VC dimension), while being rich enough to contain a reasonable approximant of the minimizer $f^\star$ of the population risk $\mathcal{R}$. As quickly outlined, in a framework stipulating in addition that the random variables Y and $f(X)$ are sub-Gaussian (see Definition A.1), the ERM is proved to yield rules with good generalization properties (see e.g. Györfi et al. [2002]; Bartlett, Bousquet, and Mendelson [2005]; Lecué and Mendelson [2016]; Massart [2007]; Boucheron, Lugosi, and Massart [2013]; Tsybakov and Zaiats [2009]; and Wainwright [2019], for a rigorous treatment of non-parametric estimation with an emphasis on convergence results similar as those introduced above; and Hastie, Tibshirani, and Friedman [2008]; Bishop [2006], for a quick overview of the different standard methods in machine learning; a purely statistical approach to statistical learning as well as in-depth treatment of the various proofs can be found in Wasserman [2004]; Devroye, Györfi, and Lugosi [1996], which treat rigorously the theoretical aspects of machine learning. Alternatively, one can refer to Proposition A.2.). Note, however, that in heavy-tail situations, alternative strategies are preferred (refer to Lugosi and Mendelson [2016], for instance).

Unfortunately those results do not apply to our survival setting, as the oberved data is not (X_i, Y_i), and need to be adapted to the censored setting if one wants to reuse the already existing tools from the machine learning literature. As most of the work in machine learning has focused on the ERM setting, and developed numerous practical tools for it, we will show how it can not only be straightforwardly adapted to deal with censored observations but more importantly we will give tail bounds on the excess risk similar to those practitioners have come to expect.[54]

54: "More importantly" mainly because practitioners are already doing empirical risk minimization on censored data, this is not novel they, however, do it without any strong theoretical justifications which is what we will solve in this chapter so that hopefully they can sleep soundly.

2.2 Risk Minimization under Censoring

About this Section

The rest of this chapter is in large part reproduced from the paper "Empirical Risk Minimization under Random Censorship" with Stéphan

Clémençon and François Portier, under final review at JMLR at the time this manuscript was written. Some of the results established in this chapter have been preliminarily presented in an elementary form at the 2018 NeurIPS Machine Learning for Healthcare workshop (ML4HEALTH) as most of the problems pertaining to credit risk have direct analogues in the medical setting.

Guillaume Ausset, Stéphan Clémençon, and François Portier (2021a). "Empirical Risk Minimization under Random Censorship". *under revision in Journal of Machine Learning Research*. arXiv: 1906.01908

As we have seen in the previous section, a wide variety of common problems can be written as the distribution-free regression problem of eq. (2.3), and time-to-event regression is no exception. In many applications such as industrial reliability (see Mann, Schafer, and Singpurwalla [1974]) or clinical trials, the random variable of interest Y to be predicted is a duration, i.e. a positive random variable, representing a *time-to-event* such as the lifespan of a manufactured component or the time to recovery of a patient.[55] In the majority of cases, the data at disposal to learn a predictive rule in the survival analysis setting is not composed of independent realizations (X_i, Y_i) of distribution $\mathbb{P}$ on the probability space $(\Omega, \mathcal{A}, \mathbb{P})$ but of observations (X_i, T_i, δ_i), where the observed durations are of the form

$$\begin{aligned} T_i &= \min(C_i, Y_i)\,, \\ \delta_i &= \mathbb{1}_{Y_i \leq C_i}. \end{aligned}$$

55: This is the *glass half-empty* formulation of survival analysis. Medical researchers are usually much more pessimistic and see it in term of time-to-death.

The random variables C_i, called here *censoring events*, model a possible right censoring while the δ_i, called here *censoring indicators*, are binary variables indicating whether censoring has occurred for each duration. The random variables C_i are used to represent any source of possible censoring that either invalidates our data or cause missing data. In the medical setting this can represent anything from someone dropping out of a medical study (the *censoring event* is therefore the last consultation at which the patient was observed alive, anything after is unknown), to something as trivial as the arrow of time (the last observed date, and therefore *censoring event*, is necessarily *today*, or at least the date of the end of the study or data collection). Of course, other types of censoring (e.g. left/interval/progressive censoring) can be encountered in practice and result in partially observed durations. Since the results established in this chapter can be straightforwardly extended to a more general framework, focus is here on the right censoring case which, though simple, covers many situations. In practice, the censoring indicator represents prior

knowledge about the sampling process and empirical performance can be greatly improved by incorporating as much knowledge as possible by taking into account as many forms of censoring as possible. Whereas the asymptotic theory of statistical estimation based on censored data is very well documented in the literature (see e.g. Fleming and Harrington [1991]; Andersen et al. [1993], and the references therein), the issues raised by censoring in statistical learning has received much less attention. It is therefore desirable to adapt the methodology presented earlier to the right censored case.

For simplification, we will assume in the rest of the chapter that the loss of interest is the squared error loss $\mathcal{L}(y, x) = \|y - x\|_2^2$ and the resulting population risk of interest the conditional regression problem

$$\mathcal{R}(f) = \mathbb{E}\left[\left(Y - f(X)\right)^2\right]. \tag{2.10}$$

Following the process of §2.1, we would normally form the empirical version of eq. (2.10) given by

$$\mathcal{R}_n(f) = \sum_{i=1}^{n} \left(Y_i - f(X_i)\right)^2, \tag{2.11}$$

which we would then minimize over some function class $\mathcal{F}$. However, as pointed earlier, we do not observe Y_i, the true duration of interest, directly but only some tangential information about it. As the empirical risk of eq. (2.11) cannot be computed directly from the data available, we rely on expressing it in terms of known quantities, or at least estimable quantities. Discarding all the censored observations to evaluate the risk of a candidate function f would lead to the quantity

$$\frac{\sum_{i=1}^{n} \delta_i \left(T_i - f(X_i)\right)^2}{\sum_{i=1}^{n} \delta_i}, \tag{2.12}$$

with $0/0 = 0$ by convention, which is clearly a biased estimate of the population risk $\mathcal{R}(f)$ in general, since, by virtue of the strong law of large numbers, it converges to $\mathbb{E}\left[\left(Y - f(X)\right)^2 \mid Y \leq C\right]$ with probability one. One may easily check that the minimizer of this functional is given by

$$\frac{\mathbb{E}\left[Y \mathbb{1}_{Y \leq C} \mid X\right]}{\mathbb{P}\left(Y \leq C \mid X\right)},$$

which significantly differs from $f^*(X) = \mathbb{E}[Y \mid X]$ in general. Similarly, taking all the observations without any correction would lead to minimizing

$$\sum_{i=1}^{n} \left(T_i - f(X_i)\right)^2, \tag{2.13}$$

with minimizer $\mathbb{E}[T \mid X]$, another biased estimate of the desired quantity.

If we assume the following hypothesis, which we will do from now on;

Assumption 2.1 (Conditional independence)**.** The random variables Y and C are conditionally independent given the input X and we have $Y \neq C$ with probability one.

We can observe that, by means of a straightforward conditioning argument, one can rewrite the population risk as

$$\mathcal{R}(f) = \mathbb{E}\left[\frac{\delta(T - f(X))^2}{S_C(T- \mid X)}\right], \tag{2.14}$$

where $S_C(u \mid x) = \mathbb{P}(C > u \mid X = x)$ denotes the conditional survival function of the random right censoring given X. We propose to estimate the risk eq. (2.14) by computing first a nonparametric estimator $\hat{S}_{C,n}(u \mid x)$ of $S_C(u \mid x)$ and then subsequently plugging it into eq. (2.14), so as to obtain

$$\widehat{\mathcal{R}}_n(f) = \frac{1}{n}\sum_{i=1}^{n} \frac{\delta_i(T_i - f(X_i))^2}{\hat{S}_{C,n}(T_i- \mid X_i)}, \tag{2.15}$$

which approximates the unknown quantity whose expectation is equal to eq. (2.14):

$$\frac{1}{n}\sum_{i=1}^{n} \frac{\delta_i(T_i - f(X_i))^2}{S_C(T_i- \mid X_i)}, \tag{2.16}$$

with the conditional survival function of C given X being unknown itself. Assumption 2.1 does not prove to be overly restrictive and is usually referred in the more general *missing data* literature (D. B. Rubin [1976]) as missing at random (MAR), as opposed to missing completely at random (MCAR) and missing not at random (MNAR). In this missing data setting we assume that the reasons that explain the censoring are fully contained in the explanatory variables available to us. This is of course, in general, a very strong hypothesis but given enough available data it is often possible to verify that is approximately true in practice. Addressing the problem in a more complex probabilistic framework, where for instance Y and C are not conditionally independent given X anymore, will be the subject of future research. The assumption stipulating that $\{Y = C\}$ is a zero-probability event is quite general, insofar as it allows considering situations where Y and/or C are discrete variables. Under conditional independence, it is obviously satisfied when the random variable Y is continuous. Note that even without Assumption 2.1, it is possible to obtain an IPCW formulation of the population risk of the form

$$\mathcal{R}(f) = \mathbb{E}\left[\frac{\delta(T - f(X))^2}{G(T-, X)}\right],$$
$$G(t, x) = \mathbb{P}\left(C > t \mid T = t,\ X = x\right).$$

Observe that the risk estimate of eq. (2.15) can be viewed as a *weighted version* of the sum of the observed squared errors of the form

$$\sum_{i=1}^{n} w_i \left(T_i - f(X_i)\right)^2, \tag{2.17}$$

just like eq. (2.12) except that the i-th weight w_i is not $\delta_i / \sum_{j \leq n} \delta_j$ anymore but

$$w_i = \frac{\delta_i}{\hat{S}_{C,n}(T_i - \mid X_i)}.$$

In the terminology of survival analysis, the weighted empirical risk of eq. (2.15) is usually referred to as an IPCW estimate (Gerds et al. [2017]) as the weighting scheme consists in attributing to the non-censored observations a weight equal to the inverse of their probability of being censored. More generally, the same principle can be found in the setting where δ indicates membership to a set such that $\delta_i = \mathbb{1}_A(x_i)$ under the name of *propensity score*, in this case the method is then called *inverse propensity score weighting* and enjoys a wide corpus of research and success in the statistical literature. The common problem meant to be corrected by propensity scores is the problem of non-uniform sampling of two or more populations. Imagine for example the scenario where one wants to study the effectiveness of some invasive treatment A in relation to a placebo B, or the absence of treatment. It is common to not be able to design a random trial and only have access to a retrospective observations. Given that the treatment A is invasive one would imagine that the only people receiving it are those with very negative survival prospects. If one were to compare the survival of the population A with that of population B naively without taking into account that the assignment depends on the covariates of the individuals, then the probable conclusion would be that treating patients kill them. It is therefore necessary to take into account the decisions and processes that lead to people falling into each sub-populations, in this case A or B and in our case censored or not censored. While not the subject of this thesis, it is interesting to see how censoring can be seen as a propensity problem in order to draw inspiration from the sampling literature.

A natural strategy to learn a predictive function in the censored framework described above then consists in solving the reweighted minimization problem

$$\inf_{f \in \mathcal{F}} \widetilde{\mathcal{R}}_n(f), \tag{2.18}$$

over an appropriate class $\mathcal{F}$. Of course the approach as described may feel circular as we have traded a known function evaluated at unknown points for an unknown function evaluated at known points. We can, of course, estimate S_C but given the symmetry in the roles of C and Y, one could

reasonably argue that we can therefore also directly estimate S_Y and use that as a plugin estimator, for example of the form

$$\hat{r}_n(x) = \int y \, \mathrm{d}\hat{\mathrm{S}}_{Y,n}(y \mid X = x),$$

for the conditional mean. We argue here that if the object of interest can be framed in the risk minimization framework then this is the problem that should be solved and the previous equation should be avoided. In the previous approach, all the error is contained in $\hat{\mathrm{S}}_{Y,n}$, which we can indeed control, but we do not have any guarantees on the quality of the subsequent plugin estimate $\hat{r}_n(x)$; a modest error in $\hat{\mathrm{S}}_{Y,n}$ can be amplified after integration. If instead the problem is framed as a risk minimization, we still have the same error in $\hat{\mathrm{S}}_{C,n}$ but we now minimize the new reweighed problem, with the effect of regaining control over the final error. As we will show later, even using unsophisticated and "wrong" $\hat{\mathrm{S}}_{C,n}$ is counterbalanced by the gains in $\hat{f}_n$. The IPCW approach proposed here is not novel and is already widely used by practitioners. Results on the properties of the IPCW risk also exist (see Gerds et al. [2017], for a recent review) but our goal here is to adapt the usual guarantees people expect under the form of tail bounds, in order to justify the current use as well as convince the rest of the survival community of the potential gains.

2.2.1 Related Work

In the rest this chapter, we will first start by building a biased plug-in estimator of the risk of eq. (2.3) by means of the Beran estimator of the conditional survival function of the censoring variable (Beran [1981]). This estimator has been studied in a similar non-asymptotic and nonparametric setting to the one considered here in the work of Dabrowska (1989), where she derives a Dvoretzky-Kiefer-Wolfowitz type exponential bound (Dvoretzky, Kiefer, and Wolfowitz [1956]) on the tail distribution of the survival function in the random design case. Similar results are obtained by van Keilegom and Veraverbeke (1996) in the fixed design case, but also extended to the quantiles. The resulting risk function of eq. (2.15), which we will refer from now on and completely arbitrarily as *IPCW risk*.[56] The asymptotic behaviour of such weighted empirical risk has been first considered in the seminal contributions of Stute and J.-L. Wang (1993); Stute (1993a,b, 1995a,b, 1996) where the convergence in distribution of the Kaplan-Meier integrals defined as

$$\int \varphi(t, x) \, \mathrm{d}F_n(t \mid x) \stackrel{\text{def}}{=} \frac{1}{n} \sum_{i=1}^{n} \frac{\delta_i \varphi(T_i \mid X_i)}{\hat{\mathrm{S}}_{C,n}(T_i)},$$

56: *Kaplan-Meier risk* will sometime be used to refer to version with unconditional estimates of the censoring weights and *Beran risk* for the conditional version when a distinction between the two cases needs to be made explicitly.

to $\int \varphi(t,x)\,\mathrm{d}F(t \mid x)$ is established. The population risk we consider is, of course, a particular instance of a Kaplan-Meier integral but not only do we wish to use weights conditional on the covariate but we are interested in non-asymptotic[57] results. These results have recently been refined in Lopez (2011); Lopez, Patilea, and van Keilegom (2013) where the complete conditional estimator of the survival is used to define the weights and convergence of the resulting new *Beran integral* is established. While these results motivate our approach as they indeed prove the convergence of our IPCW estimator of the risk to the population risk, the results obtained are still asymptotic and therefore not entirely satisfactory given our requirements. In Rotnitzky and Robins (1992) or Section 3.3 in van der Laan and Robins (2003), a parametric estimate (see D. R. Cox [1972]) of $S_C(Y \mid X)$ is used as a plugin in order to obtain an estimator of the risk. However such an approach is limited by well-known misspecification issues related to the choice of the parametric family. In order to mitigate this shortcoming, van der Laan and Robins (2003); D. Rubin and van der Laan (2007), make use of a *doubly robust loss* which allows for misspecification of either of S_C or S_T as long as one of them is correctly specified.[58] The doubly robust approach exploits the symmetry between C and Y in order to use all the observations by adding a correction term to the IPCW risk

57: As well as *uniform* in the sense that we want results for the *worse* possible case in x, that is in the supremum case.

58: Neither have to be correctly specified on all the space, it is sufficient if they are *globally* well specified jointly, potentially on disjoint spaces.

$$\frac{1}{n}\sum_{i=1}^{n}\Bigg(\underbrace{\frac{\delta_i(T_i - f(X_i))^2}{\hat{S}_{C,n}(T_i- \mid X_i)}}_{\text{IPCW}} + \underbrace{\frac{(1-\delta_i)\hat{Q}_n(T_i, X_i)}{\hat{S}_{C,n}(T_i- \mid X_i)} - \int_0^{T_i} \frac{\hat{Q}_n(u, X_i)}{\hat{S}_{C,n}(u- \mid X_i)}\,\mathrm{d}\hat{\Lambda}_{C,n}(u \mid X_i)}_{\text{Robustness Correction}}\Bigg),$$

where

$$\hat{Q}_n(y,x) = -\frac{\int_y^\infty \mathcal{L}(u, f(x))\,\mathrm{d}\hat{S}_{Y,n}(u \mid x)}{\hat{S}_{Y,n}(y \mid x)}.$$

Not only such doubly robust risk has the advantage of being correct if either Y or C is correctly specified, but the resulting estimator is locally efficient if both models are correctly specified. We constrain ourselves here to the simpler IPCW loss only for practical reasons as the proofs would be greatly complicated, but extending the results to the doubly robust case will be the focus of future research. The doubly robust approach has been further investigated in Molinaro, Dudoit, and van der Laan (2004); Steingrimsson, Diao, Molinaro, et al. (2016); Steingrimsson, Diao, and Strawderman (2019) where different methodologies are proposed to build classification trees based on the previous loss. The use of the Kaplan-Meier estimate (Kaplan and Meier [1958]) for $S_C(u \mid x)$ has been considered in several papers

(Stute [1993b, 1996]; Bang and Tsiatis [2002]; Kohler, Máthé, and Pintér [2002]). In those approaches, even if the censoring model is free from any parametric modelling, the assumptions required to ensure consistency are quite strong as the distribution of C is supposed to be independent from X (see Stute [1996], for more details). In particular, the weights used are independent from X. To overcome the previous restrictions, the Beran estimate (Beran [1981]), which is a kernel smoothing version of the Kaplan-Meier estimate,[59] can be employed instead of the unconditional Kaplan-Meier estimator or a conditional but parametric approach such as a Cox estimate. Such an approach is promoted and studied in Lopez (2011); Lopez, Patilea, and van Keilegom (2013). Based on accuracy results for kernel-based Beran estimators of the conditional survival function $S_C(\cdot \mid x)$ such as those subsequently presented, the performance of solutions of eq. (2.15) is investigated in the next section. We point out that, as highlighted in §2.5, alternative inference strategies for conditional survival function estimation can be considered but, for simplicity, we restrict our attention to kernel-smoothing techniques, although the analysis carried out can be extended to other nonparametric methods (e.g. partition-based techniques such as survival trees, nearest neighbours and more).

59: This will be introduced in greater details in the following section.

The results presented in this chapter are therefore comparable to the ones of Lopez (2011) as both are concerned with the IPCW risk, relaxing in particular the restrictive assumption on the dependence between C and X done in Stute (1993b, 1996). In Lopez (2011), an asymptotic representation of the estimation error is established when the input variable is univariate ($d = 1$). An extension with a single index model is considered in Lopez, Patilea, and van Keilegom (2013), that is when the ambient space is in general of dimension $d > 1$ but the solution lies on a subspace of dimension $s = 1$. The proof technique used in the next chapters is based on the asymptotic equicontinuity of the empirical process and imposes strong conditions on the bandwidth choice, e.g. $nh^3 \to \infty$ (see Theorem 3.3 in Lopez [2011]; Lopez, Patilea, and van Keilegom [2013], and Theorem 3.1 in), but follows the general scheme described earlier to obtain exponential tail bounds. The major difference, and difficulty, from the proofs obtained in the completely observed setting comes from the presence of an estimated quantity both in the numerator and the denominator, which amongst other consequences breaks the representation as a sum of independent and identically distributed variables. The nonasymptotic analysis carried out in this chapter is therefore quite different and relies on two crucial steps:

1. Linearization of the estimator of the risk.
2. Use of concentration results for generalized U-processes to describe its behaviour (see e.g. Clémençon and Portier [2018]).

Additionally, the approach adopted here to establish nonasymptotic rate

bounds requires weaker conditions, only that $nh^{2d}/|\log(h^d)| \to \infty$ in the d-dimensional case. We will prove that, under appropriate conditions, minimizers of the IPCW risk proposed have good generalization properties, achieving learning rate bounds of order $\sqrt{\log(n)/n}$ when ignoring the impact of model bias on the plug-in estimation step, the same as ERM in absence of any censoring. Beyond this theoretical analysis, illustrative numerical results are also provided in §2.5, providing strong empirical evidence of the relevance of the approach promoted. They reveal in particular that, even if the estimator of the conditional survival function plugged is only moderately accurate, IPCW risk minimizers significantly outperform approaches that ignore censoring, which is to be expected, but also approaches that first model the distribution and then deduce the statistics of interest (such as the conditional mean) instead of directly solving the corresponding ERM problem (i.e. the minimization of the squared error in the case of the conditional mean).

Incidentally, we point out that the problem under study can be viewed as a very specific type of *transfer learning* problem (see e.g. Pan and Q. Yang [2010]), insofar as, due to the censoring, the distribution of the training/source data is not that of the test/target data. However, the source domain coincides here with the target one and the predictive task remains the same. The same concept can usually also be found under the name *covariate shift*, where it has received considerable attention in general as it became clear to practitioners that training data did not always look like the data found in the wild[60] or more insidious that the distribution of the data can shift gradually over time.[61] Learning bounds for this type of problem, fundamentally treated as reweighted problems, exist for example in Cortes, Mansour, and Mohri (2010) but do not take into account the fact that you normally would have to estimate the weights themselves, instead considering the weights to be known.[62]

60: As it happens, it appears that the average person does not look like a 1024×1024 celebrity in photos.

61: Imagine for example a dataset of loans: it is expected that salaries will change over time because of inflation. The distribution of salaries will therefore be entirely shifted in 10 years compared to today, even for the *same* population

62: Imagine for example training on a dataset of images skewed toward one sex when you know that the general population is equidistributed: the weights are known in this case.

2.2.2 Integration domain

In order to properly introduce our results, and in order to fairly compare them to the existing literature, we start by reducing the domain of integration of the population risk. As any conditional survival function, $S_C(y \mid x)$ vanishes as y tends to infinity, it is desirable to avoid dealing with the asymptotic behaviour of the conditional survival function of the censoring and stipulating assumptions on the rate of decay of the tail. This will also deal simultaneously with the related problem of identifiability when the support of C extends past the support of Y: there is no hope of uniquely characterizing the tail of something you cannot observe.[63] Therefore, in the analysis carried out in §2.4 we restrict the study of the prediction problem to a borelian domain $\mathbb{K} \subset \mathbb{R}_+ \times \mathbb{R}^d$ such that $S_C(y \mid x)$

63: Many of the papers cited earlier sidestep this problem by specifying a model for C

stays bounded away from 0 on it and consider the IPCW risk

$$\widetilde{\mathcal{R}}_{\mathbb{K}}(f) = \mathbb{E}\left[\frac{\delta\left(T - f(X)\right)^2}{S_C(T- \mid X)}\mathbb{1}_{\mathbb{K}}\left(T, X\right)\right], \tag{2.19}$$

as well as its Beran empirical counterpart

$$\widetilde{\mathcal{R}}_{n,\mathbb{K}}(f) = \frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i(T_i - f(X_i))^2}{\widehat{S}_{C,n}(T_i- \mid X_i)}\mathbb{1}_{\mathbb{K}}\left(T_i, X_i\right). \tag{2.20}$$

For the sake of readability, from now on, we drop the explicit dependence on $\mathbb{K}$ in the notation but all the risks will implicitly be considered to be the restricted risk introduced above.

2.2.3 Outline of this chapter

As discussed at length above, the present analysis distinguishes itself from previous works, relying on the IPCW approach as well, in several aspects. First, the problem of regression in presence of censoring is tackled here from the angle of prediction and not as the problem of estimating the conditional expectation f^* with minimum $L_2(\mu_X)$-error when denoting the marginal distribution of X by μ_X. Although a Beran estimator of the survival function of C given X is involved in the empirical risk construction given above, the goal pursued here is to ensure that the predictor $\tilde{f}_n$ obtained by solving eq. (2.15) has a small *excess risk* $\mathcal{R}(\tilde{f}_n) - \mathcal{R}(f^*)$ with large probability. As will be discussed in detail in the next section, establishing nonasymptotic guarantees for statistical learning in the censored context, in the form of generalization bounds, yields technical difficulties which are far from straightforward to overcome when avoiding the simplifying assumption, hardly met in practice, that the output variable Y is independent from the random variable C modelling the censoring mechanism (see Assumption 2.1, for a much more realistic framework for regression of censored training data). In contrast, in this chapter, we derive sound theoretical results providing nonasymptotic guarantees for the minimizers of the risk by jointly estimating the intrinsic loss and the censoring mechanism.

The rest of the chapter is organized as follows:

1. The framework we consider for statistical learning based on censored training data is detailed in §2.3, where notions pertaining to survival data analysis involved in the subsequent study are also briefly recalled and a nonasymptotic uniform bound for the Beran estimator of the conditional survival function of the censoring is also stated.

2. In §2.4, the statistical version of the expected quadratic risk we propose, based on the Beran estimator previously studied, is introduced and the performance of its minimizers is analysed.
3. General illustrative numerical results are presented in §2.5 while experiments relative to finance specifically will be presented separately in chapter 5.
4. Several concluding remarks are collected in §2.7 while the proofs are delayed to §2.8 in order to not hurt readability because of their technicality and length.

2.3 Preliminary Results

In this section, we first describe at length the probabilistic setup considered for the remainder of the chapter and recall basic concepts of *censored data analysis* on which the subsequent analysis relies. Next, we establish a nonasymptotic bound for the deviation between the conditional survival function of the random censoring and its Beran estimator under adequate smoothness assumptions. Here and throughout, the indicator function of any event $\mathcal{E}$ is denoted by $\mathbb{1}_{\mathcal{E}}$ and the Dirac mass at any point x by δ_x. When well-defined, the convolution product between two real-valued Borelian functions $g(x)$ and $w(x)$ on $\mathbb{R}^d$, is denoted by

$$(g * w)(x) = \int_{\mathbb{R}^d} g(x - x')w(x')\,\mathrm{d}x'.$$

The left-limit at $s > 0$ of any càdlàg function S on $\mathbb{R}_+$ is denoted by $S(s-) = \lim_{t \uparrow s} S(t)$.

2.3.1 Kernel Estimate of the Survival

We briefly recall the Beran approach for the estimation of a conditional survival function by means of a kernel smoothing procedure and state a uniform bound for the deviations between the conditional survival function of C given X and its Beran estimator. The Beran estimator of S_C will later be a key quantity of our distribution-free framework. As shall be discussed below, this result refines those obtained in Dabrowska (1989) and Du and Akritas (2002), which are of a similar nature, except that they are related to the estimation of the conditional survival function of the duration Y given X, denoted by $S_Y(u \mid x) = \mathbb{P}(Y > u \mid X = x)$, rather than that of the conditional survival function of the censoring C given X. Define the conditional integrated hazard function of the right censoring

C given X by

$$\Lambda_C(u \mid x) = -\int_0^u \frac{S_C(\mathrm{d}s \mid x)}{S_C(s- \mid x)}. \tag{2.21}$$

and the conditional subsurvival functions

$$H(u \mid x) = \mathbb{P}(Y > u \mid X = x),$$
$$H_0(u \mid x) = \mathbb{P}(Y > u, \delta = 0 \mid X = x),$$

for $u \geq 0$ and $x \in \mathbb{R}^d$. As we have (under Assumption 2.1),

$$H_0(\mathrm{d}u \mid x) = S_Y(u- \mid x) S_C(\mathrm{d}u \mid x),$$
$$H(u- \mid x) = S_Y(u- \mid x) S_C(u- \mid x),$$

we obtain

$$\Lambda_C(u \mid x) = -\int_0^u \frac{H_0(\mathrm{d}s \mid x)}{H(s- \mid x)}.$$

Here, we propose to build an estimate of $\Lambda_C(u \mid x)$ by plugging into eq. (2.21) Nadaraya-Watson type kernel estimates of the conditional subsurvival functions and derive from it an estimator of $S_C(u \mid x)$. Of course, alternative estimation techniques can be considered for this purpose. Throughout the chapter, $K : \mathbb{R}^d \mapsto \mathbb{R}_+$ is a symmetric bounded *kernel function*, i.e. a bounded nonnegative Borelian function, integrable w.r.t. Lebesgue measure such that $\int K(x)\,\mathrm{d}x = 1$, $K(x) = K(-x)$ for all $x \in \mathbb{R}^d$ (see Wand and Jones [1994]). We assume it lies in the linear span of functions w, whose subgraphs $\{(s,u) : w(s) \geq u\}$, can be represented as a finite number of Boolean operations among sets of the form $\{(s,u) : p(s,u) \geq \zeta(u)\}$, where p is a polynomial on $\mathbb{R}^d \times \mathbb{R}$ and ζ an arbitrary real-valued function. This assumption guarantees that the collection of functions

$$\left\{ K\left(\frac{x - \cdot}{h}\right) : x \in \mathbb{R}^d,\ h > 0 \right\},$$

is a bounded VC type class (see Giné, Koltchinskii, and Zinn [2004]), a property that will be useful to establish our results. Although very technical at first glance, this hypothesis is very general and is satisfied by kernels of the form $K(x) = \zeta(p(x))$, p being any polynomial and ζ any bounded real function of bounded variation see Nolan and D. Pollard (1987) or when the graph of K is a pyramid (truncated or not). For any bandwidth $h > 0$ and $x \in \mathbb{R}^d$, we define the rescaled kernel

$$K_h(x) \stackrel{\text{def}}{=} \frac{1}{h^d} K\left(\frac{x}{h}\right).$$

Based on the kernel estimators given by

$$\hat{H}_{0,n}(u,x) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - X_i)\mathbb{1}_{T_i>u,\ \delta_i=0}, \tag{2.22}$$

$$\hat{H}_{n}(u,x) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - X_i)\mathbb{1}_{T_i>u}, \tag{2.23}$$

$$\hat{g}_{n}(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - X_i), \tag{2.24}$$

define the conditional subsurvival functions estimates

$$\hat{H}_{0,n}(u \mid x) = \frac{\hat{H}_{0,n}(u,x)}{\hat{g}_n(x)},$$

$$\hat{H}_{n}(u \mid x) = \frac{\hat{H}_{n}(u,x)}{\hat{g}_n(x)},$$

as well as the biased estimators of $\Lambda_C(u \mid x)$ and $S_C(u \mid x)$

$$\hat{\Lambda}_{C,n}(u \mid x) = -\int_0^u \frac{\hat{H}_{0,n}(\mathrm{d}s \mid x)}{\hat{H}_n(s- \mid x)}, \tag{2.25}$$

$$\hat{S}_{C,n}(u \mid x) = \prod_{s\le u}\left(1 - \mathrm{d}\hat{\Lambda}_{C,n}(s \mid x)\right), \tag{2.26}$$

with $\mathrm{d}\Lambda(t) = \Lambda(t) - \Lambda(t-)$, which are classically referred to as the conditional Nelson-Aalen and Kaplan-Meier estimators (Dabrowska [1989]).

2.3.2 Bound on the Error of the Estimate of the Survival

Let $b > 0$ and define the set

$$\gamma_b = \left\{(y,x) \in \mathbb{R}_+ \times \mathbb{R}^d : S_Y(y|x) \wedge S_C(y|x) \wedge g(x) > b\right\},$$

which is supposed to be non-empty. On this set, one may guarantee that $\hat{H}_{0,n}(y,x)$ and $\hat{H}_{0,n}(y,x)$ are both away from 0 with high probability, which permits the study of the fluctuations of eq. (2.26). In addition, the mild and standard smoothness assumption below is required in the analysis to control the estimation bias.

Assumption 2.2 (Smoothness)**.** For all $u \in \mathbb{R}_+$, the functions $x \mapsto H(u \mid x)$, $x \mapsto H_0(u \mid x)$ and $x \mapsto g(x)$ are twice continuously differentiable on $\mathbb{R}^d$ with all partial derivatives bounded by L.

The result stated below provides a uniform bound for the deviation between $S_C(u \mid x)$ and its estimator eq. (2.26).

Theorem 2.6 (Uniform bounds on the survival function)**.** *Suppose that Assumptions 2.1 and 2.2 are fulfilled. Then, there exist constants $M_1 > 0$, $M_2 > 0$ and $h_0 > 0$ depending on b, L and K only such that, for all $\varepsilon \in (0,1)$, we have with probability greater than $1-\varepsilon$:*

$$\sup_{(t,x)\in\gamma_b} \left|\hat{S}_{C,n}(t \mid x) - S_C(t \mid x)\right| \leq M_1 \left\{ \sqrt{\frac{\left|\log(h^{d/2}\varepsilon)\right|}{nh^d}} + h^2 \right\},$$

as soon as $h \leq h_0$ and $nh^d \geq M_2 \left|\log(h^{d/2}\varepsilon)\right|$.

The technical proof is given later in §2.8 (refer to the latter for a description of the constants M_1, M_2 and h_0 involved in the result stated above). A similar result, for the conditional survival function of Y given X, is proved in Theorem 2.1 of Dabrowska (1989). Observe also that choosing $h = h_n \sim n^{-1/(d+4)}$ yields a rate bound of order $\sqrt{\log(n)/n^{4/(d+4)}}$ with high probability.

Finally, we emphasize that estimation of the conditional expectation or density is not the goal we pursue here, the regression framework considered in the next section having to do with *prediction*, i.e. the construction of a predictive rule $\tilde{f}_n(X)$ from censored training data with "good" predictive capacity. Although the learning procedure we investigate in this chapter consists in minimizing a plug-in estimator of the risk of eq. (2.16) and consequently involves the nonparametric estimator of eq. (2.26), the accuracy of the prediction is measured by the excess risk, not by the estimation error $\mathbb{E}\left[(\tilde{f}_n(X) - f^*(X))^2\right]$. In addition, we point out that alternative flexible local averaging methods such as k-NN, decision trees or random forests, could naturally be used to compute estimators of $H_0(u,x)$, $H(u,x)$ and $g(x)$ and consequently estimators of $S_C(u \mid x)$ and $\Lambda_C(u \mid x)$. However, whereas the accuracy of nonparametric estimators based on kernel smoothing under appropriate smoothness hypotheses can be studied rather easily, it is much less convenient to establish rates for estimators produced by tree-based techniques for example (one generally prefers to investigate estimators built by means of variants, involving completely random splitting for instance, which are quite different from the algorithms used in practice). For this reason, the predictive performance of extensions of the statistical learning approach studied, based on estimators of $S_C(t \mid x)$ built by means of tree-based or k-NN techniques, are studied from an empirical angle only in this chapter, in §2.5.

2.4 Generalization Bounds for IPCW Risk Minimizers

It is the purpose of this section to investigate the excess of risk of eq. (2.19) related to a domain $\mathbb{K} \subset \mathbb{R}_+ \times \mathbb{R}^d$ of minimizers $\widetilde{f}_n(x)$ of the IPCW risk of eq. (2.20) over a class $\mathcal{F}$ of predictive functions that is of controlled complexity (see the technical assumptions below), while being rich enough to yield a small bias $\mathcal{R}(f^*) - \mathcal{R}(\bar{f}^*)$. We consider here the situation where, for all $i \in \{1, \ldots, n\}$, the estimate of the quantity $S_C(T_i \mid X_i)$ plugged into eq. (2.16) is obtained by evaluating the kernel smoothing estimator of $S_C(y \mid x)$ investigated in §2.3 and based on the subsample $\mathcal{D}_n^{(i)}$ at $(y, x) = (T_i, X_i)$ defined as

$$\mathcal{D}_n^{(i)} \stackrel{\text{def}}{=} \left\{ \left(X_j, T_j, \delta_j \right) : 1 \leq j \leq n, j \neq i \right\}.$$

The corresponding versions of the kernel estimators eqs. (2.22) to (2.24) and those of eqs. (2.25) and (2.26) are respectively denoted by $\hat{H}_{0,n}^{(i)}(y \mid x)$, $\hat{H}_n^{(i)}(y \mid x)$, $\hat{g}_n^{(i)}(x)$, $\hat{\Lambda}_{C,n}^{(i)}(y \mid x)$ and $\hat{S}_{C,n}^{(i)}(y \mid x)$. This yields the *leave-one-out* estimator of the risk of any candidate f

$$\widetilde{\mathcal{R}}_n(f) = \frac{1}{n} \sum_{i=1}^{n} \frac{\delta_i \left(T_i - f(X_i) \right)^2}{\hat{S}_{C,n}^{(i)} \left(T_i - \mid X_i \right)} \mathbb{1}_{(T_i, X_i) \in \mathbb{K}}, \tag{2.27}$$

that is well defined on the event $\bigcap_{i=1}^{n} \left\{ \hat{S}_{C,n}^{(i)}(T_i - \mid X_i) > 0 \right\}$. As we clearly have

$$\mathcal{R}(\widetilde{f}_n) - \inf_{f \in \mathcal{F}} \mathcal{R}(f) \leq 2 \sup_{f \in \mathcal{F}} \left| \widetilde{\mathcal{R}}_n(f) - \mathcal{R}(f) \right|,$$

the key of the analysis is the control of the fluctuations of the process $\{\widetilde{\mathcal{R}}_n(f) - \mathcal{R}(f) : f \in \mathcal{F}\}$. Slightly more generally, we establish below a uniform deviation bound for processes of type

$$Z_n(\varphi) = \left(\frac{1}{n} \sum_{i=1}^{n} \frac{\delta_i \varphi(T_i, X_i)}{\hat{S}_{C,n}^{(i)}(T_i - \mid X_i)} \right) - \mathbb{E}\left[\varphi(Y, X) \right], \ \varphi \in \Phi,$$

where the indexing class Φ fulfils the following property allowing us to control the fluctuations of the pseudo-variables $\hat{S}_{C,n}^{(i)}(T_i - \mid X_i)$, as in Theorem 2.6.

Assumption 2.3 (Restricted loss)**.** There exists a domain $\mathbb{K} \subset \gamma_b$ such that $\varphi(y, x) = 0$ as soon as $(y, x) \notin \mathbb{K}$ for all $\varphi \in \Phi$.

Equipped with these notations, when $\varphi(Y, X) = \left(Y - f(X) \right)^2 \mathbb{1}_{(T,X) \in \mathbb{K}}$ observe that

$$Z_n(\varphi) = \widetilde{\mathcal{R}}_n(f) - \mathcal{R}(f).$$

That is, as long as we restrict our definition of the loss, the risk we study theoretically matches exactly with the usual definition of the risk.

2.4.1 Linearization of the Risk

Whereas in the standard regression framework or in classification ERM can be straightforwardly studied by means of maximal deviation inequalities for empirical processes, the form of the process $\{Z_n(\varphi) : \varphi \in \Phi\}$ of interest is very complex since the terms averaged in eq. (2.20) are obviously far from being independent due to the presence of the plugged leave-one-out estimators of the quantities $S_C(T_i - \mid X_i)$. The subsequent analysis is all the more technically difficult that, in contrast to most works devoted to statistical censored data analysis, the simplifying assumption, unrealistic in many situations in practice, that Y and C are independent is avoided here, cf. Assumption 2.1. Our approach to the study of the fluctuations of the process Z_n consists in linearizing the statistic $Z_n(\varphi)$, i.e. approximating $Z_n(\varphi)$ by a standard i.i.d. average in the L_2-sense, as stated in the next proposition. In order to make this decomposition explicit, further notations are needed. We set, for all $i \in \{1, \ldots, n\}$,

$$\begin{aligned}
\hat{a}_n^{(i)}(t \mid x) &= -\int_0^t \frac{c(u \mid x)}{H(u,x)} \left(\hat{H}_{0,n}^{(i)}(\mathrm{d}u, x) - H_0(\mathrm{d}u, x)\right) \\
&\quad + \int_0^t \frac{c(u \mid x)}{H(u,x)^2} \left(\hat{H}_n^{(i)}(u,x) - H(u,x)\right) \hat{H}_{0,n}^{(i)}(\mathrm{d}u, x), \\
\hat{b}_n^{(i)}(t \mid x) &= -\int_0^t \frac{c(u \mid x)}{H(u,x)^2 \hat{H}_n^{(i)}(u,x)} \left(\hat{H}_n^{(i)}(u,x) - H(u,x)\right)^2 \hat{H}_{0,n}^{(i)}(\mathrm{d}u, x) \\
&\quad - \int_0^t \frac{\hat{S}_{C,n}^{(i)}(u- \mid x) - S_C(u- \mid x)}{S_C(u \mid x)} \hat{\Delta}_n^{(i)}(\mathrm{d}u|x),
\end{aligned}$$

where

$$\begin{aligned}
\hat{\Delta}_n^{(i)}(\mathrm{d}u|x) &= \hat{\Lambda}_{C,n}^{(i)}(\mathrm{d}u|x) - \Lambda_C(\mathrm{d}u|x), \\
c(u \mid x) &= \frac{S_C(u- \mid x)}{S_C(u \mid x)}.
\end{aligned}$$

Equipped with these notations, we can now state the following result.

Proposition 2.7 (Decomposition of the IPCW risk)**.** *Suppose that Assumptions 2.1 to 2.3 are fulfilled. There exist constants $h_0 > 0$ and $M_1 > 0$ that depends on b, L and K only such that*

(i) for any $n \geq 2$ and $\varepsilon \in (0,1)$, provided that $h \leq h_0$ and $nh^d \geq M_1 |\log(h^{d/2}\varepsilon)|$, the event

$$\mathcal{E}_n \stackrel{\text{def}}{=} \bigcap_{i \leq n} \left\{ \forall (t,x) \in \mathbb{K}, \hat{S}_{C,n}^{(i)}(t,x) \geq \frac{b}{2} \text{ and } \hat{H}_n^{(i)}(t,x) \geq \frac{b^3}{2} \right\},$$

occurs with probability greater than $1 - \varepsilon$;

(ii) for all $\varphi \in \Phi$ and $n \geq 2$, we have on the event $\mathcal{E}_n$:

$$Z_n(\varphi) = L_n(\varphi) + M_n(\varphi) + R_n(\varphi),$$

where

$$L_n(\varphi) = \frac{1}{n}\sum_{i=1}^{n}\left(\delta_i \frac{\varphi(T_i, X_i)}{S_C(T_i \mid X_i)} - \mathbb{E}\left[\delta \frac{\varphi(T, X)}{S_C(T \mid X)}\right]\right),$$

$$M_n(\varphi) = -\frac{1}{n}\sum_{i=1}^{n}\delta_i \varphi(T_i, X_i)\frac{\hat{a}_n^{(i)}(T_i \mid X_i)}{S_C(T_i \mid X_i)},$$

$$R_n(\varphi) = \frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i \varphi(T_i, X_i)}{S_C(T_i \mid X_i)}\left(-\hat{b}_n^{(i)}(T_i \mid X_i) + \frac{\left(S_C(T_i \mid X_i) - \hat{S}_{C,n}^{(i)}(T_i \mid X_i)\right)^2}{S_C(T_i \mid X_i)\hat{S}_{C,n}^{(i)}(T_i \mid X_i)}\right).$$

The proof is given in §2.8. Observe that the term $L_n(\varphi)$ is a basic centred i.i.d. sample mean statistic and its uniform rate of convergence $1/\sqrt{n}$ can be recovered by applying maximal deviation bounds for empirical processes under classic complexity assumptions such as those stipulated below, whereas the term $M_n(\varphi)$ is more complicated, since it involves multiple sums. It is dealt with by means of results pertaining to the theory of U-processes (Peña and Giné [1999]), by showing that it can be decomposed as $M_n(\varphi) = L'_n(\varphi) + R'_n(\varphi)$, the sum of a linear term and a second-order term. The term $R_n(\varphi) + R'_n(\varphi)$ is a remainder term (second order) and shall be proved to be negligible with respect to $L_n(\varphi) + L'_n(\varphi)$.

The theory of U-processes is used next to describe the uniform behaviour of $M_n + R_n$. Such concentration results are also used in Clémençon, Lugosi, and Vayatis (2008) and Papa, Bellet, and Clémençon (2016) in simpler situations, where the residuals take the form of a degenerate U-statistic. In our case, due to the presence of a leave-one-out estimate of the survival function, the U-processes that arise do not have all their diagonal terms (e.g., the sum indexes $1 \leq i, j \leq n$ are restrained to $i \neq j$). This is of particular interest because results dealing with U-processes are in most cases stated for such sums (see Lemma 2.11 and Corollary 2.12 in the Appendix section) and, more importantly, removing diagonal terms improves the estimation accuracy by reducing the bias (see also Delyon and Portier (2016), remark 4).

2.4.2 Uniform tail bounds of the excess risk

To obtain uniform concentration inequalities over the function class Φ, it is standard (Nolan and D. Pollard [1987]; Giné and Guillou [2001]) to assume the following type of control on the complexity of the class.

Assumption 2.4 (VC class)**.** The set Φ of real-valued functions on $\mathbb{R}_+ \times \mathbb{R}^d$ is a bounded VC type of class with parameter (A, v) and constant envelope M_Φ.

The formal definition of VC classes is given in the Appendix section. By means of these assumptions, the following result, proved in the Appendix section, describes the order of magnitude of the fluctuations of the process Z_n.

Proposition 2.8 (Tail bound of the excess risk)**.** *Suppose that Assumptions 2.1 and 2.4 are fulfilled. There exist constants h_0, M_1, M_2 and M_3 that depend on (A, v), M_Φ, L, K and b only, such that, for all $n \geq 2$ and $\varepsilon \in (0, 1)$, the event*

$$\sup_{\varphi \in \Phi} |Z_n(\varphi)| \leq M_1 \left(\sqrt{\frac{\log(M_2/\varepsilon)}{n}} + \frac{|\log(\varepsilon h^{d/2})|}{nh^d} + h^2 \right),$$

occurs with probability greater than $1 - \varepsilon$ provided that $h \leq h_0$, $nh^{2d} \geq M_3 |\log(\varepsilon h^d)|$.

The risk excess probability bound stated in the following theorem shows that, remarkably, minimizers of the IPCW risk attain the same learning rate as that achieved by classic empirical risk minimizers in absence of censoring, when ignoring the model bias effect induced by the plug-in estimation step (*cf* choice of the bandwidth h).

Theorem 2.9 (Uniform control of the excess risk)**.** *Suppose that Assumptions 2.1 and 2.4 are fulfilled. There exist constants h_0, M_1, M_2 and M_3 that depend on (A, v), M_Φ, L, K and b only, such that, for all $n \geq 2$ and $\varepsilon \in (0, 1)$, the event*

$$\left| \mathcal{R}(\tilde{f}_n) - \mathcal{R}(f^\star) \right| \leq M_1 \left(\sqrt{\frac{\log(M_2/\varepsilon)}{n}} + \frac{|\log(\varepsilon h^{d/2})|}{nh^d} + h^2 \right),$$

occurs with probability greater than $1 - \varepsilon$ provided that $h \leq h_0$, $nh^{2d} \geq M_3 |\log(\varepsilon h^d)|$.

The proof is a direct application of Proposition 2.8. A similar bound for the expectation of the risk excess of minimizers of the empirical IPCW risk can be classically derived with quite similar arguments, details are left to the reader. We finally point out that, given that Proposition 2.8 holds true for a fairly general class of functions Φ, the guarantees provided by Theorem 2.9 can be naturally extended to more general risk measures than that defined by the quadratic loss.

2.5 Numerical Experiments

Beyond the theoretical generalization guarantees established in the previous section, we now examine at length the performance of the predictive approach proposed in the context of regression based on censored data from an empirical perspective. We present various experiments using both synthetic and real data, and compare it to alternative methods documented in the survival analysis literature standing as natural competitors. As shall be seen below, the experimental results obtained provide strong empirical evidence of the relevance of the Kaplan-Meier empirical risk minimization approach described in section §2.3 and analysed theoretically in section §2.4. All the experiments and figures displayed in this chapter can be reproduced using the code available at `git.sr.ht/~aussetg/locallinear`.

2.5.1 Experimental Setup

Before presenting and discussing the numerical results obtained, we first describe the experimental schemes used here to investigate the predictive capacity of the learning procedure under random censoring previously studied.

Data Generative Models In all the synthetic experiments we have carried out, the generation of the data is based either on the proportional hazard model of D. R. Cox and Oakes (1984) or else on the accelerated time failure model of Buckley and James (1979); both commonly used for parametric modelling and statistical estimation of conditional survival functions in the censored setup. Samples of the triplet (T, δ, X) are obtained by specifying the marginal distribution of X, as well as the conditional distribution of (Y, C) given X. For simplicity, the input r.v. X is here uniformly distributed on the unit square $[0,1]^d$, for $d \in \{2, 4, 8\}$. Only the results for $d = 4$ are presented below, while those obtained for $d \in \{2, 8\}$ are available through the link mentioned above.

Cox Model. The first survival model we use to simulate synthetic data stipulates that

$$\begin{aligned} S_Y(y \mid x) &= \exp\big(-\exp\big(\beta^\intercal x\big)\, y\big)\,, \\ S_C(y \mid x) &= \exp\big(-\exp\big(\beta_C^\intercal x\big)\, y\big)\,, \end{aligned} \tag{2.28}$$

where β and β_C are parameters in $\mathbb{R}^d$. Given X, the conditional distribution of Y is thus exponential with parameter $\exp(\beta^\intercal X)$, while that of C is exponential with parameter $\exp(\beta_C^\intercal X)$.

Accelerated Failure Time Model (AFT). The second generative model we considered assumes that

$$\begin{aligned} \log(Y) &= -\beta^\intercal X + \varepsilon_0, \\ \log(C) &= -\beta_C^\intercal X + \varepsilon_1, \end{aligned} \tag{2.29}$$

where the r.v. ε_0 (respectively ε_1) is independent from X. Different accelerated failure time models can thus be generated, depending on the distributions D_0 and D_1 chosen for ε_0 and ε_1. Three distributions have been used: Normal (N with mean and variance $(3/2, 1)$, Laplace (L) with location and scale $(1, 1)$ and Gamma (G) with shape and scale $(0, 1)$. Denoting by $\mathrm{AFT}(D_0, D_1)$ the model such that $(\varepsilon_0, \varepsilon_1) \sim D_0 \otimes D_1$, the variants $\mathrm{AFT}(N, N)$, $\mathrm{AFT}(N, L)$ and $\mathrm{AFT}(N, G)$ have been simulated. Since the results obtained for these AFT models are quite similar to those based on the Cox model, only the latter are presented below. We refer to the link aforementioned for a description of the results based on the data generated through the AFT models.

Parameters β and β_C. In the Cox and AFT models, the level of censoring can be tuned by carefully choosing the parameters β and β_C. In order to guarantee that the censoring is informative, we use the following parametrization:

$$\begin{aligned} \beta^\intercal &= \big[\overbrace{1 \quad \dots \quad 1}^{\lceil d/2 \rceil} \quad 0 \quad \dots \quad 0\big], \\ \beta_C^\intercal &= \lambda \big[1 \quad 0 \quad 1 \quad 0 \quad 1 \quad \dots\big], \end{aligned}$$

where the tuning parameter $\lambda > 0$ controls the level of censoring $1 - p$ with $p = \mathbb{P}(Y \leq C)$ and $u \in \mathbb{R} \mapsto \lceil u \rceil$ is the ceiling function. For a targeted censoring level p, the parameter λ can be empirically determined so that $\sum_{i=1}^n \delta_i \simeq np$.

Plugged estimator of the conditional survival function $S_C(\cdot \mid x)$ The estimate of the risk eq. (2.20) one seeks to minimize is partly determined by the choice of the estimator $\hat{S}_{C,n}(\cdot \mid x)$ of $S_C(\cdot \mid x)$ plugged into it. We consider for $\hat{S}_{C,n}$ the kernelized Kaplan-Meier estimator eq. (2.26), in its standard version denoted by $\hat{S}_{C,n}^{\mathsf{Kern}}(\cdot \mid x)$ and in its leave-one-out version as well, denoted by $\hat{S}_{C,n}^{(i)\mathsf{Kern}}(\cdot \mid x)$. We denote by $\hat{S}_{C,n}^{\mathsf{KM}}(\cdot)$ the Kaplan-Meier estimator of the unconditional survival function of C, which can be seen as the limit of $\hat{S}_{C,n}^{\mathsf{Kern}}$ when $h \to \infty$ and yields the Kaplan-Meier risk considered in Stute (1995a). In addition, we used $\hat{S}_{C,n}^{(i)\,\mathsf{KNN}}(\cdot \mid x)$, the estimator obtained by replacing the kernel smoothing involved in eq. (2.20) by a nearest neighbour averaging, in a leave-one-out fashion. Finally, we

also considered $\hat{S}^{\texttt{RF}}_{C,n}$, the survival random forest estimator proposed in Ishwaran, Kogalur, et al. (2008). From each estimator of S_C, one computes a plug-in estimation of the IPCW risk:

$$\begin{aligned}
&\text{Kernel} \quad \sum_{i=1}^{n} \delta_i \frac{(T_i - f(X_i))^2}{\hat{S}^{\texttt{Kern}}_{C,n}(T_i \mid X_i)} \qquad \text{LOO} \quad \sum_{i=1}^{n} \delta_i \frac{(T_i - f(X_i))^2}{\hat{S}^{(i)\texttt{Kern}}_{C,n}(T_i \mid X_i)} \\
&\text{Forest} \quad \sum_{i=1}^{n} \delta_i \frac{(T_i - f(X_i))^2}{\hat{S}^{\texttt{RF}}_{C,n}(T_i \mid X_i)} \qquad \text{Stute} \quad \sum_{i=1}^{n} \delta_i \frac{(T_i - f(X_i))^2}{\hat{S}^{\texttt{KM}}_{C,n}(T_i)} \qquad (2.30) \\
&k\text{-NN} \quad \sum_{i=1}^{n} \delta_i \frac{(T_i - f(X_i))^2}{\hat{S}^{(i)\texttt{KNN}}_{C,n}(T_i \mid X_i)} \qquad \text{Oracle} \quad \sum_{i=1}^{n} \delta_i \frac{(T_i - f(X_i))^2}{S_C(T_i \mid X_i)},
\end{aligned}$$

or if one allows biased estimated:

$$\text{Naive} \quad \sum_{i=1}^{n} (T_i - f(X_i))^2 \qquad \text{Observed} \quad \sum_{i=1}^{n} \delta_i \, (T_i - f(X_i))^2 \,. \qquad (2.31)$$

The naive and observed empirical risks introduced above correspond to strongly biased estimators of the population risk eq. (2.10) of course. Note that we ignore here the normalizing constant for the sake of brevity; multiplicative constant that is irrelevant if one is only after the minimizer of the ERM problem. However, if one is interested in the estimate of the true risk itself, it is then necessary to correctly normalize the previous quantities in order to obtain complete case estimators. A point of comparison is the true oracle risk

$$\text{True Oracle} \quad \frac{1}{n} \sum_{i=1}^{n} (Y_i - f(X_i))^2 \,,$$

i.e. the empirical risk in absence of any censoring (i.e. when all the Y_i's are observed). For each risk, a prediction rule $\tilde{f}_n$ is built by (approximately) minimizing it over a certain class $\mathcal{F}$. The results are depicted in fig. 2.4 for various sizes of the (censored) training sample and different censoring levels, the prediction error being evaluated by means of a test (uncensored) sample of size 5000: learning a predictive function by minimizing an IPCW estimator (here IPCW leave-one-out (LOO)) of the risk always outperforms naive alternatives, the gain in predictive performance naturally becoming more pronounced as the level of censoring $1 - p$ increases. Unsurprisingly, when most of the points are observed (i.e. $p \to 1$), all methods reach roughly the same error, all the losses in eq. (2.30) being equal for $p = 1$, as can be observed in fig. 2.5. Note in fig. 2.5 that the IPCW estimator performs the best comparatively to the naive methods for a moderate level of censoring which can be explained by the fact that when the censoring is inexistent the methods are equivalent but when the censoring is too

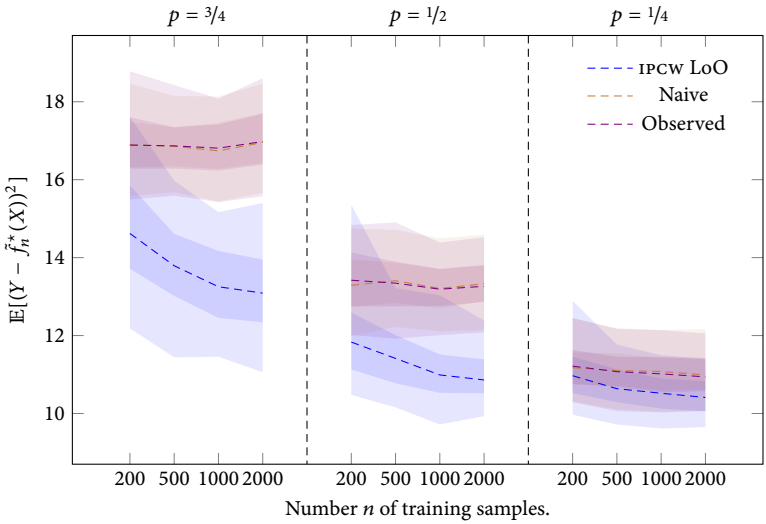


FIGURE 2.4: Prediction error for data generated by the Cox model of eq. (2.28), when minimization is performed over the class of affine predictive rules.

extreme there is not enough data to estimate the IPCW weights. This is, of course, exaggerated in the present example as the training set contains only 1000 observations which with 90% of censoring results in effectively 100 observations available for the conditional estimation of the weights.

Truncation of the estimator $\hat{S}_{C,n}(\cdot \mid x)$. In the theoretical analysis carried out in the previous section, we placed ourselves on a restricted set γ_b. However, in practice, we employ a truncation approach by simply removing the last jump of the estimated survival function. For instance, $\hat{S}^{\text{Kern}}_{C,n}(y|x)$ is taken as

$$\prod_{\substack{\tilde{Y}_i \leq y \\ \tilde{Y}_i < \max_{j:\delta_j=0} Y_j}} \left(1 - \delta \hat{\lambda}_{C,n}(T_i \mid x)\right). \tag{2.32}$$

Observe that, although it is not a survival function anymore, it is still a relevant estimator. This alleviates possible difficulties caused by the frequent edge case where the last individual is observed ($\delta = 1$), since, in the case where eq. (2.26) is used, we have then $\delta_n / \hat{S}_{C,n}(T_n \mid X_n) = \infty$. Of course, it would have been possible to decide to force a restriction on γ_b by considering the flooring $\max(b, \hat{S}_{C,n}(y \mid x))$ rather than $\hat{S}_{C,n}(y \mid x)$. However, b then becomes an hyperparameter of the procedure that has to be tuned by the practitioner. It is common in the survival analysis literature to only consider the restricted mean survival time, i.e. $\min(Y, \tau)$ as a more relevant and easier to learn metric instead of the true Y (see for example Royston and Parmar (2011) or Steingrimsson and Morrison

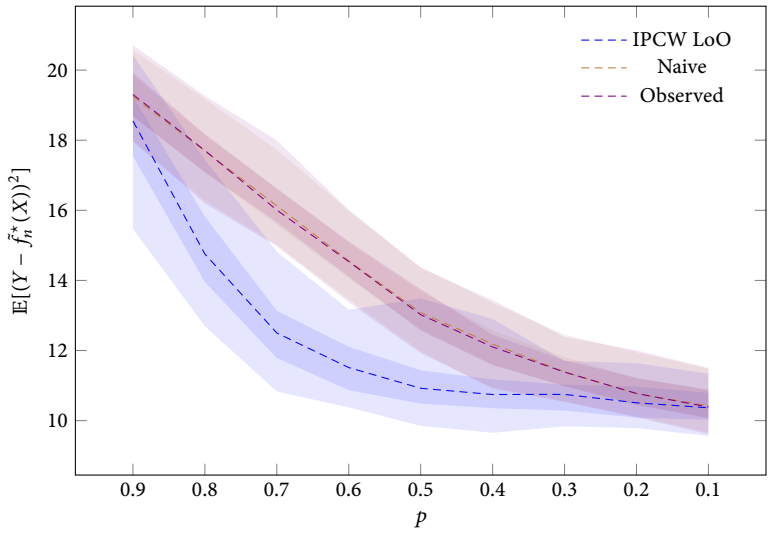


FIGURE 2.5: Prediction error for data generated by the Cox model of eq. (2.28) with $n = 1000$, when minimization is performed over the class of affine predictive rules and varying levels of censoring p.

(2020)). In this case the survival function of the restricted variable is equal to the truncated survival function of the true variable. Our approach therefore consists in using the restricted mean censoring time as the target for the weights in order to reduce the noise. We evaluated truncation of the survival function and flooring by comparing the predictive performance attained by the rule learnt from data generated by means of the Cox model, when choosing successively $\hat{S}_C^{\texttt{Kern}}$, $\hat{S}_{C,n}^{(i)\texttt{Kern}}$ and $\hat{S}_{C,n}^{(i)\texttt{KNN}}$ for $\hat{S}_{C,n}$. As depicted by fig. 2.6 for the specific case where $\hat{S}_{C,n} = \hat{S}_{C,n}^{(i)\texttt{Kern}}$ (and this remains true in the other cases), a wrong choice for b may have serious consequences, while the truncation approach of eq. (2.32) consistently produces good results. Consequently, the truncated version is always used in the following experiments.

Calibration of $\hat{S}_{C,n}(\cdot \mid x)$. In order to fully specify the estimator $\hat{S}_{C,n}(\cdot \mid x)$, it may be necessary to choose specific hyperparameters. Without censoring, and therefore without having to resort to the IPCW approach, one would select the various hyperparameters by way of a cross-validation; this approach is, however, impossible in our case as the loss itself is unknown and only estimated, worse any modification of the parameters of $\hat{S}_C$ results in a modification of the estimator of the loss we wish to minimize. One possible solution is to rely on a *surrogate loss* i.e. an auxiliary loss that we are able to compute exactly and on which a cross-validation is therefore possible. For $\hat{S}_{C,n}^{\texttt{Kern}}$ and $\hat{S}_{C,n}^{(i)\texttt{Kern}}$, we consider $\hat{m}_{h,n}(x)$ the nonparametric kernel regression of T w.r.t. X, known as the Nadaraya-Watson estimator,

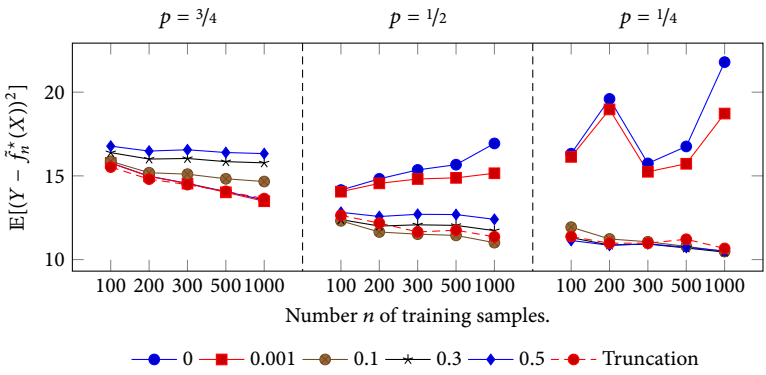


FIGURE 2.6: Prediction error $\mathbb{E}[(Y - \tilde{f}_n(X))^2]$ when $\mathcal{F}$ is the class of affine functions, choosing the IPCW LOO risk estimator and the Cox model of eq. (2.28) for generating the data.The curves correspond to different floors b and to the truncation approach of eq. (2.32).

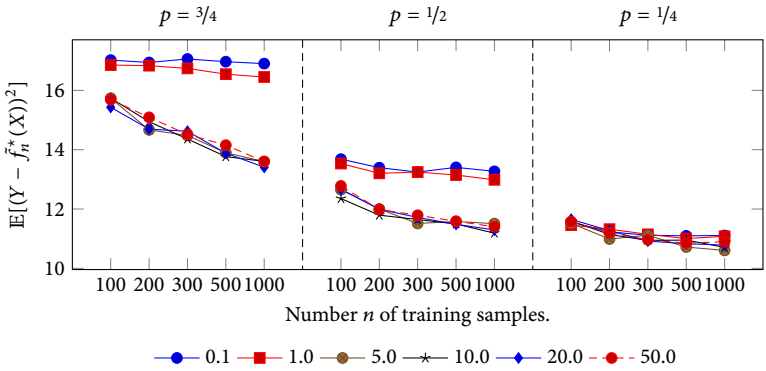


FIGURE 2.7: Prediction error $\mathbb{E}[(Y - \tilde{f}_n(X))^2]$ for varying bandwidth h, with $\mathcal{F}$ a random forest model, IPCW LOO estimator of S_C, and data following the Cox model of eq. (2.28).

and the surrogate loss $\mathbb{E}\left[|T - \hat{m}_h(X)|^2\right]$ which is then minimized using cross-validation with respect to h. In this way, a value for the bandwidth parameter h^*_{CV} is obtained and might be used in $\hat{S}^{\texttt{Kern}}_{C,n}$ and $\hat{S}^{(i)\texttt{Kern}}_{C,n}$. This approach is also easily applied to set the number of neighbours involved in $\hat{S}^{(i)\texttt{KNN}}_{C,n}$. As a close relative of the task of estimating S_C, the previous regression loss is a good candidate for the surrogate cross-validation. In the specific case of $\hat{S}^{(i)\texttt{Kern}}_{C,n}$ and $\hat{S}^{(i)\texttt{KNN}}_{C,n}$, we experimentally studied the impact of the choice of h and k on the prediction performance defined here by the L^2 loss,

$$\mathbb{E}\left[\left(Y - \tilde{f}_n(X)\right)^2\right].$$

The results for the specific case of $\hat{S}^{(i)\texttt{Kern}}_{C,n}$ are given in fig. 2.7 and demonstrate the need to be over-conservative rather than under-conservative in the choice of these hyperparameters. In our experiments, choosing h at least equal to the value h^*_{CV} obtained by minimizing the surrogate, and up to 5 times this value, is a safe choice. Consequently, we use $h = 5h^*_{\text{CV}}$ in

the following experiments. For $\hat{S}_C^{\text{RF}}$, given the large number of hyperparameters, the default parameters selected by the package's authors have been used.

2.5.2 Experimental Results based on Synthetic Data

We now present the results obtained from the data generated by means of the model previously described.

Risk estimation. While not the focus of the predictive approach studied in this chapter, it is of interest to evaluate the quality of the estimation of $\mathbb{E}[\varphi(Y, X)]$, related to a certain function φ, attained by the IPCW method. In order to make computations easier, we choose to study functionals of the form $\varphi(Y, X) = Y \exp(-X^\intercal \beta)$, where Y follows the Cox model described in eq. (2.28). In this case $\mathbb{E}[\varphi(Y, X)] = 1$. For a single random dataset $\mathcal{D}_n = \{(X_i, Y_i, \delta_i) : i = 1, \ldots, n\}$ of size n, the excess risk is given by

$$\mathcal{E}_n(\beta) = \left| 1 - \frac{1}{n} \sum_{i=1}^{n} \frac{\delta_i}{\hat{S}_{C,n}(T_i \mid X_i)} T_i \exp(-X_i^\intercal \beta) \right|.$$

Based on $M = 100$ simulated datasets, we study the distribution of $\mathcal{E}_n(\beta)$ for varying sample sizes n. We represent the median, 5% and 95% quantiles of $\mathcal{E}_n(\beta)$ in fig. 2.8 for each survival estimator. As can be seen in fig. 2.8,

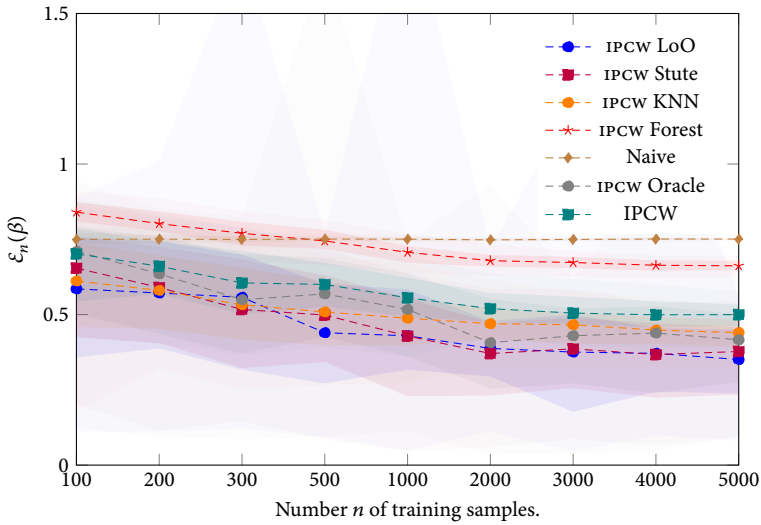


FIGURE 2.8: Estimation error for the IPCW Risks of eq. (2.30) compared to the naive method, for $p = 1/4$ and data following the Cox model of eq. (2.28).

while the naive uncorrected method results in a poor approximation of the considered expectation (as expected, since it is strongly biased), the IPCW

reweighting errors converge towards 0. One should pay specific attention to the particularly good performance of the leave-one-out version of the IPCW estimators. We also point out that low-bias estimators of S_C, such as the Random Forest estimator, can underperform significantly compared to their high bias counterparts such as the unconditional Stute estimator. This behaviour is consistent with the observations made in the previous discussion about calibration. It is illustrated by fig. 2.7. We empirically observe that the IPCW estimator of the risk with oracle weights (i.e. computed from the true conditional survival function $S_C(. \mid x)$) may be less accurate than plug-in versions (i.e. computed from an estimator of the conditional $\hat{S}_{C,n}(. \mid x)$) and exhibits a much higher variance. Intuitively, this phenomenon can be explained by the fact that the empirical weights governed by the value $1/S_C(Y_i|X_i)$ can grow arbitrarily large for observations in the tail. This phenomenon is reduced for the estimated version LOO (resp. k-NN) because of the truncation (see the implementation details above) and the over-conservative choice of the bandwidth (resp. of the number of neighbours). A similar phenomenon occurs for estimated of importance sampling type, for which the weights appearing in the denominator need to be tuned finely, see Delyon and Portier (2020).

Predictive performance. In this chapter, we are concerned with the *predictive* task, rather than risk estimation. Hence, we now focus on the problem of minimizing eq. (2.10). We study the prediction error[64] $\mathcal{R}(\tilde{f}_n)$, that is

$$\mathcal{R}(\tilde{f}_n) = \mathbb{E}\left[\left(Y - \tilde{f}_n(X)\right)^2\right],$$

for the following types of predictive model $\mathcal{F}$: Support vector Regression (SVR), Random Forests and Linear Regression. Although the choices we made are far from being exhaustive, they correspond to tools commonly used by practitioners.

Following the experimental scheme presented in §2.5.1 we first generate train sets of varying size n and test sets of fixed size 5000 according to the data generative models described in §2.5.1. We then estimate on the train set the weights corresponding to each risk described in eq. (2.30) with hyperparameters chosen using the procedure given in §2.5.1 before computing $\tilde{f}_n$, the minimizer of the resulting empirical risk over the class $\mathcal{F}$ considered. We finally estimate the prediction error $\mathcal{R}(\tilde{f}_n)$ using the test dataset. While not a perfect or exact estimate of the true prediction error, or population risk, for a sufficiently large test the estimate is sufficiently accurate for comparisons.[65] Each experiment is entirely replicated (including the sampling of the train and test sets) 100 times in order to obtain reliable statistics for the distribution of the true risk of the learning procedure.

64: Or population risk.

65: Note that in that case we only have to rely on *classical* results of convergence of the estimate as $\tilde{f}_n$ is independent of the test set.

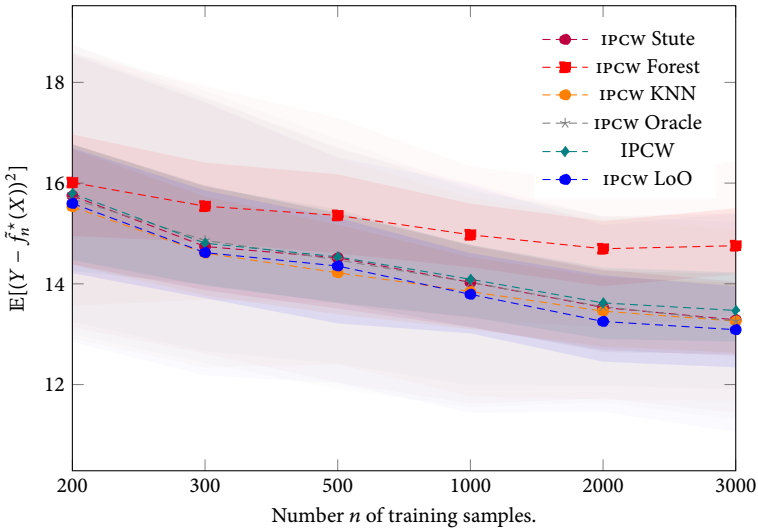


FIGURE 2.9: Prediction error for different estimators of S_C using the linear regression model for data generated by the Cox model eq. (2.28).

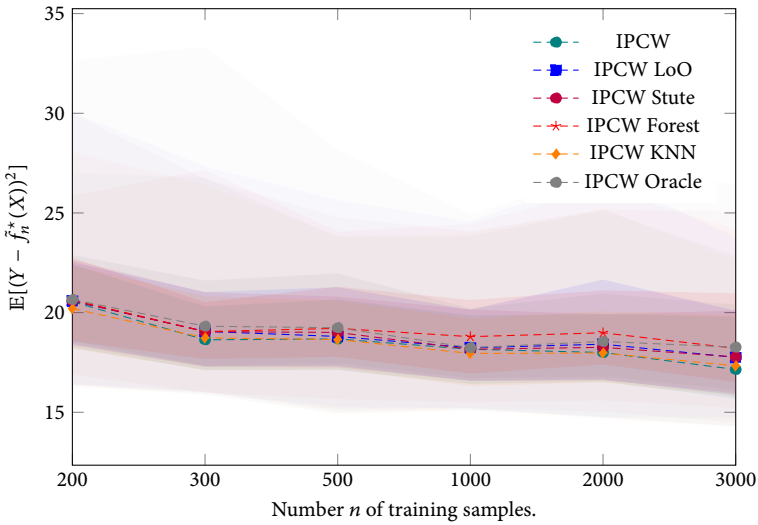


FIGURE 2.10: Prediction error for different estimators of S_C using the linear regression model for data generated by the AFT(N, L) model eq. (2.29).

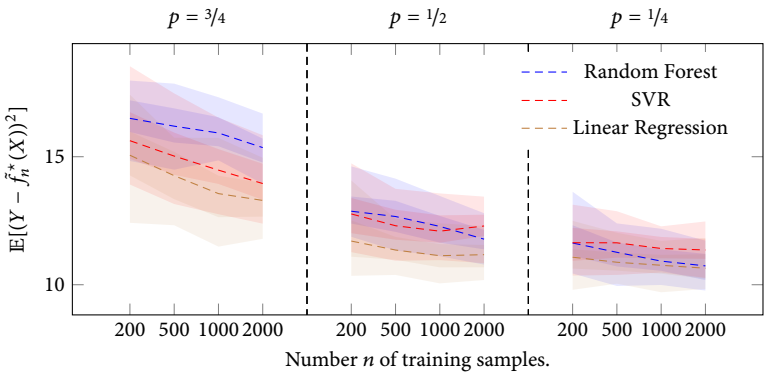


FIGURE 2.11: Prediction error for the three predictive models (SVR, random forest and linear regression) using the IPCW LOO for data generated by the Cox model of eq. (2.28).

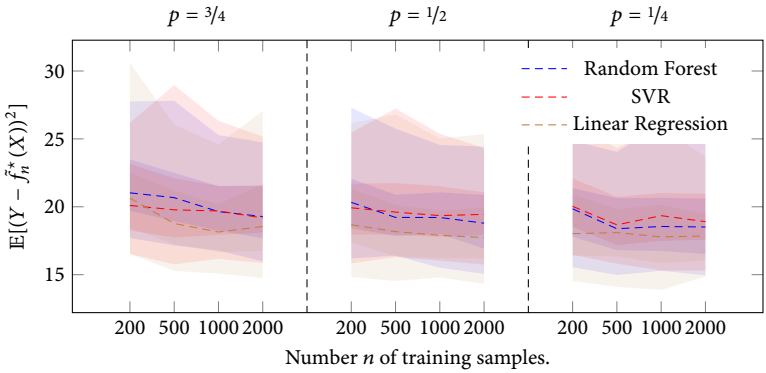


FIGURE 2.12: Prediction error for the three predictive models (SVR, random forest and linear regression) using the IPCW LoO for data generated by the AFT(N, L) model eq. (2.29).

As already observed earlier in figs. 2.8 and 2.9 shows that the IPCW LOO predictor systematically outperforms the other predictors in our experiments, no matter the level of censoring p and across different distributions as can be seen here for the specific case of AFT(N, L). Consequently, any further mention of IPCW implicitly refers to the IPCW LOO version from now on and all subsequent experiments involve the use of $\hat{S}_{C,n}^{(i)\texttt{Kern}}$. We also underline that these results hold true, no matter the predictive model $\mathcal{F}$ considered, as can be seen by examining fig. 2.11 or the underlying distribution as can be seen by examining fig. 2.12. It is interesting to note that the difference between the methods increase as the number of observations in the training set increases. As the conditional estimators are more complex, more data is necessary in order to differentiate themselves from the unconditional versions.

Finally, we compare variants of popular machine learning methods implementing the IPCW technique promoted in this chapter to standard state-of-the-art procedures from the survival analysis literature that do not rely on re-weighted risk minimization. Such techniques include classic statisti-

cal methods based on the preliminary estimation of the survival function, as already mentioned in §2.1 (see e.g. van der Laan and Robins [2003]), the survival function estimator being next used to estimate the downstream quantity of interest in a plug-in fashion, provided that the latter can be expressed as an integral w.r.t. the survival function, just like the conditional mean. An alternative approach, in the spirit of machine learning methods, consists in designing losses tailored to the censored regression problem, either through transformation models in Van Belle, Pelckmans, Suykens, and Huffel (2011), or else by adapting the SVM methodology, as done in e.g. Van Belle, Pelckmans, Suykens, and Van Huffel (2007); Pölsterl, Navab, and Katouzian (2015, 2016). We also include the method of Hothorn et al. (2006) which shares similarities with the approach investigated in this chapter and uses a boosting technique to optimize a loss reweighted by unconditional Kaplan-Meier weights, as well as the technique proposed in Ishwaran, Kogalur, et al. (2008) that builds a recursive splitting of the feature space $\mathbb{X}$ by maximizing a measure of inter-cluster dissimilarity of the survival functions, the resulting clusters being then used for downstream tasks such as classification, regression, or quantile estimation. We compare the predictive performance of ten estimators of the regression function based on statistical models documented in the survival literature with that of five predictive functions learned using the IPCW risk minimization approach. The IPCW versions of the machine learning techniques for regression considered in these experiments, corresponding to the approach studied in the present chapter, have been implemented with `Scikit-Learn` (Pedregosa et al. [2011]), combined with our own implementation of the LOO IPCW predictor we propose. For the survival machine learning methods mentioned above, we use the reference implementations of the `Scikit Survival` package (Pölsterl [2020]). The canonical implementation of Ishwaran and Kogalur (2007) is used for Random Survival Forest. The default values for the hyperparameters are used in every case. All experiments are based on $n = 200$ training observations. Results for all methods can be found in table 2.1. While the undeniable superiority of IPCW methods compared to the standard survival techniques may appear surprising at first glance. However, keeping in mind that the performance measure is here the prediction error $\mathcal{R}(\hat{f}_n)$, or expected squared error, it is expected that directly minimizing an estimator of the prediction error yields better results than a two-stage procedure that consists in estimating first the underlying distribution and forming next an estimator of its mean.

2.5.3 Experimental Results based on Real Data

The performance of the IPCW risk minimization approach is now investigated on the The Cancer Genome Atlas (TCGA) Cancer data (Grossman

Table 2.1: Performance on the Cox dataset

	Method	$\mathbb{E}\left[\left(Y - \tilde{f}_n(X)\right)^2\right]$ $p = 3/4$	$p = 1/2$	$p = 1/4$
Scikit Survival	Survival Gradient Boosting	3.19	3.55	3.61
	Cox Proportional Hazards	7.86	7.61	7.03
	Coxnet	7.62	7.39	6.85
	Kernel Survival SVM	4.02	3.92	4.13
	Survival SVM	4.04	4.09	3.94
	Hinge Loss Survival SVM	8.10	8.28	8.09
	Minlip Survival SVM	3.27	3.96	4.22
	Random Survival Forest	2.01	2.94	2.78
Scikit Learn	Ridge + IPCW	**1.75**	**1.49**	**1.24**
	Kernel Ridge + IPCW	2.07	1.60	1.35
	Linear Regression + IPCW	1.81	**1.49**	**1.24**
	Random Forest + IPCW	1.85	1.57	1.36
	SVR + IPCW	1.87	1.66	1.42

et al. [2016]) using solely the ribonucleic acid (RNA) transcriptomes as informative variables. All models are trained on $n = 8080$ patients with a censoring rate of 18%, we measure on the remaining 1449 observed patients the prediction error, as well as the concordance index defined by

$$C(f) = \frac{\sum_i \sum_j \delta_j \mathbb{1}_{f(X_j) > f(X_i)} \mathbb{1}_{T_j \leq T_i}}{\sum_i \sum_j \delta_j \mathbb{1}_{T_j \leq T_i}}, \tag{2.33}$$

which can be seen as an extension of the classical area-under-curve (AUC) metric for the standard classification problem to censored data, measuring how *well ordered* the predicted death times are. Note that as a complete case statistic itself, the glsauc could benefit from the same methodology as presented in this chapter; that is from the use of IPCW weights to construct an estimator of the concordance. In practice as the concordance index is not the object of importance here, we use the standard definition of Harrel in order to facilitate comparisons with other approaches as well as prevent any discussions on how to choose the specific IPCW approach to compute the weights. For all the results, we use the IPCW methodology presented earlier. The Cox proportional hazards model was, however, learned after variable selection via a Lasso regression so as to augment performance.

We observe from the results in table 2.2 that, as expected, the predictors built through IPCW risk minimization significantly outperform their competitors, including the standard Cox model, for the prediction task. The large improvement compared to the Cox approach is not unexpected, insofar as the seemingly less sophisticated IPCW approach is specifically designed for the purpose of *prediction*, but can still surprise. By directly minimizing an estimate of the loss of interest, it naturally achieves a lower test prediction error than that reached by the traditional two-stage ap-

Method	IPCW		Naive		Observed	
	L^2 Error (*years*)	C	L^2	C	L^2	C
Cox	18.78	0.6095	–	–	–	–
SVR	2.768	0.563	2.796	0.575	2.795	0.543
Lin. Reg.	3.193	0.594	4.971	0.557	3.898	0.508
Ridge	3.193	0.594	4.962	0.557	3.896	0.508
Kernel Ridge	2.683	0.597	2.704	0.592	2.956	0.513
Random Forest	**2.577**	**0.630**	2.636	0.603	2.878	0.542

TABLE 2.2: Results of the IPCW approach on the TCGA Cancer data.

proach in statistics, which consists in estimating first the distribution and deducing next an estimator of the minimizer of the loss of interest. The Cox estimator only controls the likelihood of the model without any concern for the predictive performance. In particular, extreme errors are not penalized in any way while those are hurtful to the overall L^2 error. More interestingly, we see that while the difference is not as pronounced, the IPCW predictors also outperform the Cox estimator with respect to the concordance index. The concordance index, as an extension of the Wilcoxon-Mann-Whitney or glsauc which therefore involves sorting the observations is a measure that can hardly be optimized directly in practice but is often viewed as valuable by practitioners. Recently, by framing sorting as the dual of an optimal transport problem, approaches to *differentiable sorting* have been proposed in Cuturi, Teboul, and Vert (2019), and Blondel et al. (2020) and could therefore be used to directly optimize an arbitrarily close approximation of the concordance index provided that the learning algorithm uses gradient descent. Remarkably, as the IPCW risk minimization approach can be combined with highly sophisticated learners (such as random forests), without any modification or increase in complexity, it is possible to significantly increase its predictive capacity, while edging the standard survival techniques on auxiliary metrics as well.

2.6 Joint IPCW Games

We have concentrated in the analysis as well as the examples our efforts on the task of predicting the regression function of the survival time by means of IPCW ERM, that is predicting Y by solving an ERM problem reweighted by the inverse probability of censoring. Regression, that is estimating $\mathbb{E}[Y \mid X]$ is, however, not the only quantity of interest one can express as an empirical risk minimization problem. In particular, for proper scoring rules[66] (Dawid and Musio [2014]), it is possible to write the problem of estimating S as an ERM problem. Notably, the Brier score (BS) (Brier and

66: That is loss functions, usually in classification, that are minimized uniquely by the true probability distribution. Note that it is usually not the case in classification as the optimal Bayes rule is invariant under scaling and translation of the threshold.

Allen [1951]), defined as

$$\mathrm{BS}_Y(t, \theta) = \mathbb{E}\left[(S_Y(t \mid X) - \mathbb{1}_{Y>t})^2\right], \tag{2.34}$$

admits $S(t)$ as a minimizer, as does the Binomial log-likelihood (BLL) (Kvamme, Borgan, and Scheel [2019]) defined by

$$\mathrm{BLL}_Y(t, \theta) = \mathbb{E}\left[-\log(1 - S_Y(t))\,\mathbb{1}_{Y\leq t} - \log(S_Y(t))\,\mathbb{1}_{Y>t}\right]. \tag{2.35}$$

Of course, both rules are inappropriate for the estimation of S_Y globally as they only concern themselves with estimation at a specific time t. It is, however, possible to integrate the previous quantities in order to obtain the integrated Brier score (IBS)

$$\begin{aligned}\mathrm{IBS}_Y(\theta) &= \int_0^\infty \mathrm{BS}_Y(t, \theta)\,\mathrm{d}t \\ &= \int_0^\infty \mathbb{E}\left[(S_Y(t \mid X) - \mathbb{1}_{Y>t})^2\right]\mathrm{d}t,\end{aligned}$$

and integrated Binomial log-likelihood (IBLL)

$$\begin{aligned}\mathrm{IBLL}_Y(\theta) &= \int_0^\infty \mathrm{BLL}_Y(t, \theta)\,\mathrm{d}t \\ &= \int_0^\infty \mathbb{E}\left[-\log(1 - S_Y(t))\,\mathbb{1}_{Y\leq t} - \log(S_Y(t))\,\mathbb{1}_{Y>t}\right]\mathrm{d}t.\end{aligned}$$

In the survival setting, given that both eq. (2.34) and eq. (2.35) are risk minimization problems, we can rewrite them in their IPCW form, that is

$$\mathrm{BS}_Y(t, \theta) = \mathbb{E}\left[\frac{S_Y(t \mid X)^2 \delta \mathbb{1}_{Y\leq t}}{S_C(Y- \mid X)} + \frac{(1 - S_Y(t \mid X))^2\,\mathbb{1}_{Y>t}}{S_C(t \mid X)}\right], \tag{2.36}$$

$$\begin{aligned}\mathrm{BLL}_Y(t, \theta) = \mathbb{E}\Big[&-\frac{\log(S_Y(t \mid X))\,\delta \mathbb{1}_{Y\leq t}}{S_C(Y- \mid X)} \\ &- \frac{\log(1 - S_Y(t \mid X))\,\mathbb{1}_{Y>t}}{S_C(t \mid X)}\Big].\end{aligned} \tag{2.37}$$

and corresponding integrated IPCW versions. Given the previous remarks, it is therefore possible to construct an estimator of S_Y if we are able to estimate S_C, as done previously. However, the variables C and Y have an entirely symmetrical role in the survival analysis setting studied here, it is therefore also entirely possible to write the same problem but in terms of

C instead of Y, that is

$$\begin{aligned}\mathrm{BS}_C(t,\theta) = \mathbb{E}\Big[&\frac{S_C(t \mid X)^2\,(1-\delta)\,\mathbb{1}_{C\le t}}{S_Y(C- \mid X)} \\ &+ \frac{(1-S_C(t \mid X))^2\,\mathbb{1}_{C>t}}{S_Y(t \mid X)}\Big], \\ \mathrm{BLL}_C(t,\theta) = \mathbb{E}\Big[&-\frac{\log(S_C(t \mid X))\,(1-\delta)\,\mathbb{1}_{C\le t}}{S_Y(C- \mid X)} \\ &- \frac{\log(1-S_C(t \mid X))\,\mathbb{1}_{C>t}}{S_Y(t \mid X)}\Big].\end{aligned}$$

Given that we then have one ERM problem depending on $\hat{S}_C$ to estimate $\hat{S}_Y$ and one ERM problem depending on $\hat{S}_Y$ to estimate $\hat{S}_C$, it seems natural to ask whether it is possible to solve both at the same time jointly. Goldstein et al. (2021) shows that it is indeed possible for strictly proper scoring rules such as BS and BLL: the authors propose an *inverse-weighted game* where both current estimates $\hat{S}_Y$ and $\hat{S}_C$ are jointly updated iteratively following algorithm 1 where the value functions V_Y and V_C can be constructed from the possible IPCW strictly proper scoring rules such as $\mathrm{BS}(t,\theta)$ or $\mathrm{BLL}(t,\theta)$ by defining the integrated variants as a collection of K games where K is the number of unique event times. In that case, for a strictly proper scoring rule S we define the value functions as

$$V(\theta) = \left(S(T_1,\theta), \ldots, S(T_K,\theta)\right).$$

If one defines

$$g_Y(\eta_Y) \overset{\text{def}}{=} J_h^\intercal \, \mathrm{d}V_Y(h(\eta_Y)),$$

where h is an invertible mapping from $\mathbb{R}^d$ to the space of parameters Θ.[67] Surprisingly, the previous scheme is able to be competitive or even beat competing approaches in the low data regime on a variety of real datasets, as can be seen in fig. 2.13.

67: We use h solely to make sure the resulting problem is unconstrained.

Algorithm 1 Following Gradients in Inverse-Weighted Games

Require: Choice of value functions V_Y and V_C, normalization function h, learning rate γ

1: **initialize** η_Y and η_C randomly
2: **repeat**
3: `// compute both gradients simultaneously`
4: $\eta_Y \leftarrow \gamma g_Y$ and $\eta_C \leftarrow \gamma g_C$
5: **until** convergence
6: $k^\star, \lambda^\star \leftarrow \mathrm{argmin}_{k,\lambda}\, \texttt{error}_{k,\lambda}$
7: **return** $k^\star, \lambda^\star$

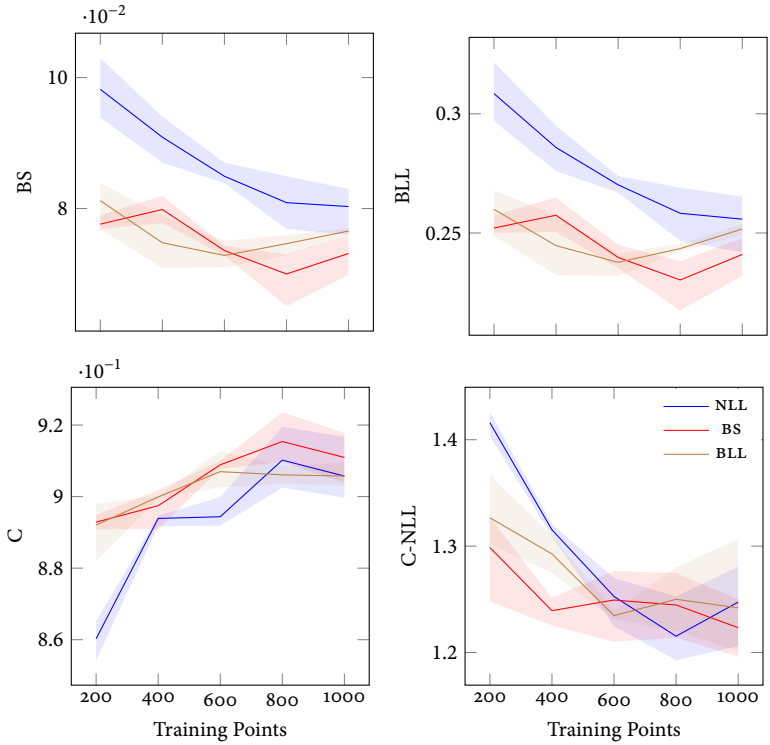


FIGURE 2.13: Performance of IPCW games for 4 metrics.

2.7 Conclusion

Machine Learning has achieved great success in the recent years by taking a different approach to the usual statistical literature. By focusing on the objectively easier problem of *prediction* instead of *estimation*, the statistical learning field has been able to frame the problem in a different way, resulting in a different approach. As seen, many prediction problems can be straightforwardly framed as minimization problems, therefore bypassing the need for the much more ambitious task of estimating the density, survival or any other quantity that define the law of the object of interest. While this approach has proven to work exceptionally well for prediction, the difference in questions asked calls for different results. We have described earlier one of the important guarantees the field of machine learning and predictive methods in general has come to expect: the ability to guarantee good generalizations properties for the models developed; that is the ability to bound the difference in error between the ideal case and the model at hand. While the bounds obtained are in reality fairly lax and are not sufficient to ensure good performances,[68] they are in practice enough to motivate the use of most methods: as long as the complexity of the function class of interest is controlled, we can expect good generalization capabilities. Such guarantees are not only incredibly useful in theory but can also lead to practical extensions like *structural risk minimization* (Vapnik and Chervonenkis [1974])[69] where the risk is minimized such that the complexity is also minimized by penalizing the overall risk term. More importantly, the empirical risk minimization framework and matching guarantees, made possible the ability to treat separately the theory and the practice: by providing broad and largely applicable results on generalizations, practitioners are able to mostly ignore the theoretical justification as long as they know their problem can be framed as an ERM problem and their proposed estimator proved to have finite VC dimension. It is therefore not surprising that the ERM framework has proven so popular.

Unfortunately, some problems cannot be written straightforwardly as sums of i.i.d. variables and therefore do not fit the ERM framework and cannot rely on the vast sum of results it entails. As seen earlier, prediction of censored observations do not fit the ERM framework as the variables are not observed. It was, however, already known, and used in practice, that the problem could be rewritten as a reweighed ERM problem and therefore solved using the tools already developed for uncensored prediction. Despite the wide use and great success of this reweighted IPCW approach, no guarantees on the generalization error existed. Even though the resulting problem looks strikingly similar to the usual ERM approach, careful examination actually reveals that it is fundamentally different: the terms in the

68: In many cases the models perform *much* better than the bounds suggest. In the case of deep learning, for example, naive VC bounds are so wide that the common assumption is that it should *not* work well. The surprisingly good results of deep-learning therefore call for better arguments than naive complexity bounds to explain their performance.

69: Теория Распознавания Образов: Статистические Проблеми Обучения (*Theory of Pattern Recognition: Statistical Problems of Learning*), only available in Russian.
Я же говорил, что советские уже все изобрели.

sum are not i.i.d. at all as the weights themselves needs to be estimated and depend on all the elements of the sum. Despite this fundamental difference, we have shown in this chapter that after linearization it was possible to obtain a problem that *looks like* the standard ERM problem, and obtained in Theorem 2.9 the usual generalization guarantee practitioners have come to expect and rely on.

The results presented here, however, are only preliminary in the sense that we only prove them for Beran type of kernel estimators of the survival function S_C. While the experiments presented in §2.5 prove that such estimators are more than enough to achieve satisfactory performances (and are often state of the art), we hint on the possibility to replace the fairly unsophisticated kernel estimator with other estimators such as RSF. Given the fact that estimation of the survival function is an important problem in itself but is now also an important ingredient for the downstream task of learning a more general regression function, it is worthwhile to devote time and efforts to the task of developing a highly accurate and flexible conditional estimator of S_C that is amenable to unstructured data such as text or even images. While it is possible to treat an unstructured data point by first transforming it into a manageable form, i.e. a vector in $\mathbb{R}^d$, by using an auxiliary model to compute its *embedding* models that are trained *end-to-end*, that is which incorporates this embedding phase directly, generally outperform significantly models that are not. Given these remarks, in the next chapter we will focus on novel neural methods for the estimation of S_C (as well as f_C), that can be applied to complex and varied datasets.

2.8 Proofs

This section contains the proofs of the results given earlier. For the sake of readability, those have been separated from the main body but the schema of the proofs is sufficiently interesting in its own right to deserve its own section instead of being relegated to appendix A.

2.8.1 Concentration Inequalities for VC Classes and Permanence Properties

For completeness, concentration results as well as preservation properties of VC classes, extensively used in the subsequent proofs, are recalled. For the sake of generality, this section is independent from the rest of the memoir. For a function $f : \mathbb{S} \mapsto \mathbb{R}$, for $\mathbb{A} \subseteq \mathbb{S}$, we define

$$\|f\|_\infty = \sup_{x \in \mathbb{S}} |f(x)|,$$
$$\|f\|_\mathbb{A} = \sup_{x \in \mathbb{A}} |f(x)|.$$

Concentration inequalities over VC classes. The following concentration inequalities provide uniform bounds on empirical sums over VC classes of functions. We start by recalling the definition of a VC class.

Definition 2.3 (VC covering)**.** Let $(\mathbb{S}, \mathcal{S})$ be a measurable space. A class $\mathcal{F}$ of real-valued functions defined on $\mathbb{S}$ is called VC of parameter $(A, v) \in]0, \infty[^2$ and constant envelope $U_{\mathcal{F}} > 0$ if for any probability measure Q on $(\mathbb{S}, \mathcal{S})$ and any $\varepsilon \in]0, 1[$:

$$\mathcal{N}(\mathcal{F}, L^2(Q), \varepsilon U_{\mathcal{F}}) \leq \left(\frac{A}{\varepsilon}\right)^v,$$

where $\mathcal{N}(\mathcal{F}, L^2(Q), \varepsilon)$ denotes the smallest number of $L^2(Q)$-balls of radius less than ε required to cover class Φ (covering number), see e.g. Nolan and D. Pollard (1987) and Giné and Guillou (2001).

The following inequality for empirical processes over VC classes is stated in Einmahl and Mason (2000); Giné and Guillou (2001) under various forms. The present version is taken from Giné and Sang (2010).

Lemma 2.10 (Uniform control of centred sums)**.** *Let $\xi_1, \xi_2, \ldots$ be i.i.d. r.v.'s valued in a measurable space $(\mathbb{S}, \mathcal{S})$ and $\mathcal{U}$ be a class of functions on $\mathbb{S}$, uniformly bounded and of VC-type with constant (A, v) and envelope $U : \mathbb{S} \to \mathbb{R}$. Set $\sigma^2(u) = \mathbb{V}[u(\xi_1)]$ for all $u \in \mathcal{U}$. There exist constants $C_1 > 0, C_2 \geq 1, C_3 > 0$ (depending on v and A) such that $\forall t > 0$, if*

$$C_1 \sigma \sqrt{n \log\left(\frac{2\|U\|_\infty}{\sigma}\right)} \leq t \leq \frac{n\sigma^2}{\|U\|_\infty}, \tag{2.38}$$

is satisfied, then

$$\mathbb{P}\left(\left|\sum_{i=1}^n (u(\xi_i) - \mathbb{E}[u(\xi_i)])\right|_{\mathcal{U}} > t\right) \leq C_2 \exp\left(-C_3 \frac{t^2}{n\sigma^2}\right).$$

The previous result is extended to the case of degenerated U-processes over VC classes (Major [2006], Theorem 2).

Lemma 2.11 (Uniform control of U-statistics)**.** *Let $\xi_1, \xi_2, \ldots$ be an i.i.d. sequence of random variables taking their values in a measurable space $(\mathbb{S}, \mathcal{S})$ and distributed according to a probability measure $\mathbb{P}$. Let $\mathcal{H}$ be a class of functions on $\mathbb{S}^k$ uniformly bounded such that $\mathcal{H}$ is of VC type with constants (A, v) and envelope G. For any $H \in \mathcal{H}$, set $\sigma^2(H) = \mathbb{V}[H(\xi_1, \ldots, \xi_k)]$ and assume that for all $j \in \{1, \ldots, k\}$, with probability 1 we have*

$$\mathbb{E}\left[H(\xi_1, \ldots, \xi_k) \mid \xi_1, \ldots, \xi_{j-1}, \xi_{j+1}, \ldots, \xi_k\right] = 0. \tag{2.39}$$

Then, there exist constants $C_1 > 0$, $C_2 \geq 1$, $C_3 > 0$ *(depending on* v *and* A*) such that for all* $t > 0$ *satisfying*

$$C_1\sigma\left(n\log\left(\frac{2\|G\|_\infty}{\sigma}\right)\right)^{k/2} \leq t \leq \sigma\left(\frac{n\sigma}{\|G\|_\infty}\right)^k, \tag{2.40}$$

then

$$\mathbb{P}\left\{\Big|\sum_{(i_1,\ldots,i_k)} H(\xi_{i_1},\ldots,\xi_{i_k})\Big|_{\mathcal{H}} > t\right\} \leq C_2\exp\left(-C_3\frac{1}{n}\left(\frac{t}{\sigma}\right)^{2/k}\right),$$

where

$$\|G\|_\infty^2 \geq \sigma^2 \geq \|\mathbb{V}[H]\|_{\mathcal{H}}^2.$$

The following result is directly derived from that stated above by specifying an appropriate value of t.

Corollary 2.12 (Uniform control of U-statistics (Alternate formulation))**.** *Let* $\xi_1, \xi_2, \ldots$ *be an i.i.d. sequence of random variables taking their values in a measurable space* $(\mathbb{S}, \mathcal{S})$ *and distributed according to a probability measure* $\mathbb{P}$*. Let* $\mathcal{H}$ *be a class of functions on* $\mathbb{S}^k$ *uniformly bounded such that* $\mathcal{H}$ *is of* VC *type with constants* (A, v) *and envelope* G*. For any* $H \in \mathcal{H}$*, set* $\sigma^2(H) = \mathbb{V}[H(\xi_1,\ldots,\xi_k)]$ *and assume that with probability* 1*, and for all* $j \in \{1,\ldots,k\}$*,*

$$\mathbb{E}\left[H(\xi_1,\ldots,\xi_k) \mid \xi_1,\ldots,\xi_{j-1},\xi_{j+1},\ldots,\xi_k\right] = 0.$$

Then, there exist constants $C_1 > 0$, $C_2 \geq 1$, $C_3 > 0$ *(depending on* v *and* A*) such that*

$$\mathbb{P}\left(\left\|\sum_{(i_1,\ldots,i_k)} H(\xi_{i_1},\ldots,\xi_{i_k})\right\|_{\mathcal{H}} \leq t(n,\sigma,\varepsilon)\right) \geq 1-\varepsilon,$$

with

$$t(n,\sigma,\varepsilon) = \sigma n^{k/2}\left(C_1\left(\log\left(\frac{2\|G\|_\infty}{\sigma}\right)\right)^{k/2} + \left(\frac{\log(C_2/\varepsilon)}{C_3}\right)^{k/2}\right),$$

provided that

$$\|G\|_\infty^2\left(C_1^{2/k}\log\left(\frac{2\|G\|_\infty}{\sigma}\right) + \frac{\log(C_2/\varepsilon)}{C_3}\right) \leq n\sigma^2,$$
$$\sup_{H\in\mathcal{H}}\sigma^2(H) \leq \sigma^2 \leq \|G\|_\infty^2.$$

VC type classes of functions - Permanence properties. In the subsequent sections, several results are obtained by applying the concentration bounds recalled above to specific classes of functions built up from the elements of the class Φ and other functions such as $K_h(x)$, $S_C(u \mid x)$ or $g(x)$. To show that these specific classes are VC, we rely on the following lemmas which exhibits situations where the VC type property is preserved, while controlling the constants (A, v) involved. In what follows the kernel K is assumed to satisfy the hypotheses introduced in §2.3.

Lemma 2.13 (see Nolan and D. Pollard (1987), Lemma 22, Assertion (ii))**.** *The class*

$$\left\{z \mapsto K\left(\frac{x-z}{h}\right) : x \in \mathbb{R}^d,\ h > 0\right\},$$

is a bounded VC class of functions.

The following result is an extension of a result established in the proof of Proposition 8 in Portier and Segers (2018).

Lemma 2.14 (Marginal VC class)**.** *Let (V, W) be a pair of random variables taking their values in $\mathbb{R}^q$ and in $\mathbb{R}^d$ respectively, denote by $f_0(v \mid W)$ the density of the conditional distribution of the random variable V given W, supposed to be absolutely continuous w.r.t. Lebesgue measure on $\mathbb{R}^q$. Let $\mathcal{F}$ be a bounded VC class of functions defined on $\mathbb{R}^q \times \mathbb{R}^d$ with parameter (A, v) and constant envelope $U_{\mathcal{F}}$. The class $\mathcal{G} = \{w \in \mathbb{R}^d \mapsto \mathbb{E}[f(V, W) \mid W = w] : f \in \mathcal{F}\}$ is a bounded VC class of functions with parameter (A, v) and constant envelope $U_{\mathcal{F}}$.*

Proof. Let Q be a probability measure on $\mathbb{R}^d$. Consider $\widetilde{Q}$ the probability measure defined through

$$\mathrm{d}\widetilde{Q}(v) = \int f_0(v \mid w)\, Q(\mathrm{d}w)\, \mathrm{d}v.$$

Let $\varepsilon \in (0, 1)$ and consider the centres $f_1, \ldots, f_N$ of an $\varepsilon U_{\mathcal{F}}$-covering of the VC class $\mathcal{F}$ with respect to the metric $L_2(\widetilde{Q})$. Let $g \in \mathcal{G}$, i.e., $g : w \in \mathbb{R}^d \mapsto \mathbb{E}[f(V, W) \mid W = w]$ with f in $\mathcal{F}$. Define $g_k = \mathbb{E}[f_k(V, W) \mid W = w]$, for $k = 1, \ldots, N$. There exists $k \in \{1, \ldots, N\}$ such that

$$\begin{aligned}
&\int \left(g(w) - g_k(w)\right)^2\ Q(\mathrm{d}w) \\
&\quad \le \int \mathbb{E}\left[\left(f(V, W) - f_k(V, W)\right)^2 \mid W = w\right]\ Q(\mathrm{d}w) \\
&\quad = \iint \left(f(v, w) - f_k(v, w)\right)^2 f_0(v \mid w)\, \mathrm{d}v Q(\mathrm{d}w) \\
&\quad = \int \left(f(v, w) - f_k(v, w)\right)^2\ \widetilde{Q}(\mathrm{d}v) \\
&\quad \le \varepsilon^2 U_{\mathcal{F}}^2,
\end{aligned}$$

using Jensen's inequality and Fubini's theorem. Consequently, we have:

$$\mathcal{N}\left(\mathcal{G}, L^2(Q), \varepsilon U_{\mathcal{F}}\right) \leq \mathcal{N}\left(\mathcal{F}, L_2(\bar{Q}), \varepsilon U_{\mathcal{F}}\right) \leq \left(\frac{A}{\varepsilon}\right)^{v}.$$

Since the constant $U_{\mathcal{F}}$ is an envelope for the class $\mathcal{G}$, the result is established. □

Lemma 2.15 (Kernel integral VC class)**.** *Let Ψ be a VC class of functions defined on $\mathbb{R}^q \times \mathbb{R}^d$ with constant envelope $U > 0$ that satisfies the following Lipschitz property: for all $\psi \in \Psi$, $z \in \mathbb{R}^q$, $(x, y) \in \mathbb{R}^d \times \mathbb{R}^d$,*

$$|\psi(z, x) - \psi(z, y)| \leq \kappa \|x - y\|,$$

with $\kappa > 0$. Let $K : \mathbb{R}^d \to \mathbb{R}$ be a positive function such that $\int K(u)\,\mathrm{d}u = 1$ and $v_K = \int \|u\|^2 K(u)\,\mathrm{d}u < \infty$. The class $\mathcal{F} = \{(z, x) \mapsto \int \psi(z, x - hu)K(u)\,\mathrm{d}u : \psi \in \Psi, 0 < h \leq \tilde{h}\}$ is a bounded measurable VC class of functions with constant envelope $(\kappa\tilde{h}\sqrt{v_K} + U)$.

Proof. Let $0 < \varepsilon \leq 1$ and $h_k = k\varepsilon\tilde{h}$, $k = 1, \ldots, \lfloor 1/\varepsilon \rfloor$, an $(\varepsilon\tilde{h})$-subdivision of the interval $]0, \tilde{h}]$. Let Q be a probability measure on $\mathbb{R}^q \times \mathbb{R}^d$. For each k, define μ_k as the probability measure of the random variable $(Z, X - h_k U)$ when $(Z, X, U) \sim Q \times \mathcal{K}$. Let $\Psi_{k,j}$, $j = 1, \ldots, N$ be an εU-cover of the function class Ψ with respect to $L^2(\mu_k)$. Let $h \in]0, \tilde{h}]$ and $\psi \in \Psi$. For any measurable function f and any k, we have

$$\left\| \int f(z, x - h_k u)K(u)\,\mathrm{d}u \right\|_{L^2(Q)} \leq \|f\|_{L^2(\mu_k)}.$$

As a consequence, for each k there exists j such that

$$\left\| \int (\psi(z, x - h_k u) - \psi_{k,j}(z, x - h_k u))K(u)\,\mathrm{d}u \right\|_{L^2(Q)} \leq \varepsilon U,$$

Besides, from the Lipschitz property, there exists k such that

$$\left\| \int (\psi(z, x - hu) - \psi(z, x - h_k u))K(u)\,\mathrm{d}u \right\|_{L^2(Q)} \leq \varepsilon\kappa\tilde{h}\sqrt{v_K},$$

The triangle inequality allows claiming that there exists j and k such that

$$\left\| \int (\psi(z, x - hu) - \psi_{k,j}(z, x - h_k u))K(u)\,\mathrm{d}u \right\|_{L^2(Q)} = \varepsilon(\kappa\tilde{h}\sqrt{v_K} + U).$$

There are $1/\varepsilon \times A\varepsilon^{-v}$ such functions $\Psi_{k,j}$ meaning that

$$\mathcal{N}\left(\mathcal{F}, \|\cdot\|_{L^2(Q)}, \varepsilon(\kappa\tilde{h}\sqrt{v_K} + U)\right) \leq A\varepsilon^{-(v+1)},$$

where $(\kappa\tilde{h}\sqrt{v_K} + U)$ is indeed an envelope for the class $\mathcal{F}$. □

We conclude the section by a preservation result for the product and the inverse.

Lemma 2.16 (Permanence of VC class)**.** *Suppose that $\mathcal{F}$ and $\mathcal{G}$ are two VC classes defined on $\mathbb{S}$ with parameters $(A_\mathcal{F}, v_\mathcal{F})$ and $(A_\mathcal{G}, v_\mathcal{G})$ and constant envelopes $U_\mathcal{F}$ and $U_\mathcal{G}$, respectively. Then it holds:*

(i) The class $\mathcal{FG} = \{fg : f \in \mathcal{F},\ g \in \mathcal{G}\}$ is VC with parameter $(2(A_\mathcal{F} \vee A_\mathcal{G}), v_\mathcal{F} + v_\mathcal{G})$ and envelope $U_\mathcal{F}U_\mathcal{G}$.

(ii) In addition, if for all $f \in \mathcal{F}$ and $x \in \mathbb{S}$, $f(x) \geq b_\mathcal{F}$, then the class $\mathcal{F}^{-1} = \{1/f : f \in \mathcal{F},\ g \in \mathcal{G}\}$ is VC with parameters $(A_\mathcal{F}U_\mathcal{F}/b_\mathcal{F}, v_\mathcal{F})$ and envelope $1/b_\mathcal{F}$.

Proof. Let $0 < \varepsilon \leq 1$ and f_k, $k = 1, \ldots, N_\mathcal{F}$ the centres of an $(\varepsilon U_\mathcal{F})$-covering of $\mathcal{F}$. Similarly, denote by g_k, $k = 1, \ldots, N_\mathcal{G}$ the centres of an $(\varepsilon U_\mathcal{G})$-covering of $\mathcal{G}$. By applying the operation $(f_k \wedge U_\mathcal{F}) \vee (-U_\mathcal{F})$, we can assume without loss of generality that the f_k (resp. g_k) are bounded by $U_\mathcal{F}$ (resp. $U_\mathcal{G}$). Then for any $f \in \mathcal{F}$ and $g \in \mathcal{G}$, there are $k \in \{1, \ldots, N_\mathcal{F}\}$ and $j \in \{1, \ldots, N_\mathcal{G}\}$ such that

$$\left\|fg - f_k g_j\right\| \leq U_\mathcal{G} \left\|f - f_k\right\| + U_\mathcal{F} \left\|g - g_j\right\| \leq 2\varepsilon U_\mathcal{G} U_\mathcal{F},$$

which implies that

$$\mathcal{N}\left(\mathcal{FG}, L_2(Q), 2\varepsilon U_\mathcal{G} U_\mathcal{F}\right) \leq \left(\frac{A_\mathcal{F}}{\varepsilon}\right)^{v_\mathcal{F}} \left(\frac{A_\mathcal{G}}{\varepsilon}\right)^{v_\mathcal{G}}$$

Taking $\varepsilon' = 2\varepsilon$ gives the result. For the second point, taking $f_k \geq b_\mathcal{F}$, we have

$$\left\|f^{-1} - f_k^{-1}\right\| \leq \frac{1}{b_\mathcal{F}^2} \left\|f - f_k\right\| \leq \frac{U_\mathcal{F}}{b_\mathcal{F}^2}\varepsilon,$$

and the result follows taking $\varepsilon' = (U_\mathcal{F}/b_\mathcal{F})\varepsilon$. □

2.8.2 Integration results

In this section we establish useful bounds related to these quantities: kernel smoothers, integrals with respect to signed measures, survival functions and hazard functions namely. This corresponds to Lemmas 2.17 to 2.20, respectively. As the previous section, this section is independent from the rest of the paper.

Lemma 2.17 (Kernel approximation error bound)**.** *Let ω an open convex subset of $\mathbb{R}^d$. Suppose that f is twice differentiable on ω such that the greatest eigenvalue of the Hessian matrix is uniformly bounded by $M > 0$, then, if the kernel K is symmetric, i.e., $K(u) = K(-u)$, we have: for all $h > 0$,*

$$\sup_{x \in \omega} |(K_h * f)(x) - f(x)| \leq \frac{M}{2} h^2 \int \|z\|^2 K(z)\, \mathrm{d}z. \tag{2.41}$$

Proof. The proof follows the same lines as the proof of Lemma 11 given in Delyon and Portier (2020). □

Lemma 2.18 (Dudley's lemma). *Let $\theta \in (0,1)$, $h : \mathbb{R}_+ \to [1, \infty[$ be Borelian, increasing, with limit $1/\theta$ at $+\infty$ and ν be any signed measure on $\mathbb{R}_+$. Then, we have: $\forall T > 0$, $\forall t \in [0,T]$,*

$$\left|\int_0^t h \, \mathrm{d}\nu\right| \leq \frac{2}{\theta} \sup_{s\in[0,T]} \left|\int_0^s \mathrm{d}\nu\right|.$$

Proof. Recall first the identity

$$\sup_{t\geq 0}\left|\int_0^t \mathrm{d}\nu\right| = \sup_{f\in \mathrm{DE}}\left|\int f \, \mathrm{d}\nu\right|, \tag{2.42}$$

where DE is the space of non-increasing functions valued in $[0,1]$ and vanishing at infinity (see e.g. Dudley [1992]). Since h is increasing from 1 to $1/\theta$, we have for any signed measure ν (whose restriction to $[0,T]$ is denoted by $\nu_{[0,T]}$),

$$\begin{aligned}\left|\int_0^t h \, \mathrm{d}\nu\right| &= \frac{1}{\theta}\left|\int_0^t \mathrm{d}\nu + \theta\int_0^t (h - \theta^{-1}) \, \mathrm{d}\nu\right| \\ &\leq \frac{2}{\theta}\sup_{f\in\mathrm{DE}}\left|\int f \, \mathrm{d}\nu_{[0,T]}\right|.\end{aligned}$$

Then applying eq. (2.42) we obtain that

$$\begin{aligned}\left|\int_0^t h \, \mathrm{d}\nu\right| &\leq \frac{2}{\theta}\sup_{s\geq 0}\left|\int_0^s \mathrm{d}\nu_{[0,T]}\right| \\ &= \frac{2}{\theta}\sup_{s\in[0,T]}\left|\int_0^s \mathrm{d}\nu\right|.\end{aligned}$$

□

Lemma 2.19 (Survival difference's bound). *Let $\tau > 0$. Let $S^{(1)}$ and $S^{(2)}$ be càdlàg non-increasing functions on $\mathbb{R}_+$ such that $S^{(1)}(0) = S^{(2)}(0) = 1$ and $S^{(2)}(\tau) \geq \theta > 0$. For $k \in \{1,2\}$, with corresponding cumulative hazard function*

$$\lambda^{(k)}(t) = -\int_0^t \frac{S^{(k)}(\mathrm{d}u)}{S^{(k)}(u-)},$$

We have:

$$\left\|S^{(1)} - S^{(2)}\right\|_{[0,\tau]} \leq \frac{2}{\theta}\left\|\Lambda^{(1)} - \Lambda^{(2)}\right\|_{[0,\tau]}.$$

Proof. Let $t \in [0,\tau]$. As $S^{(2)}(t) > 0$, the integration by part argument of Fleming and Harrington (1991), Theorem 3.2.3 yields

$$\begin{aligned}\frac{S^{(1)}(t) - S^{(2)}(t)}{S^{(2)}(t)} &= -\int_0^t \frac{S^{(1)}(u-)}{S^{(2)}(u)} \left(\Lambda^{(1)}(\mathrm{d}u) - \Lambda^{(2)}(\mathrm{d}u)\right) \\ &= -\int_0^t S^{(1)}(u-)\Delta_1(\mathrm{d}u), \end{aligned} \tag{2.43}$$

where we set

$$\Delta_1(\mathrm{d}u) = \frac{\Lambda^{(1)}(\mathrm{d}u) - \Lambda^{(2)}(\mathrm{d}u)}{S^{(2)}(u)}.$$

We then apply the integration by parts formula (refer to Shorack and Wellner [2009], page 305, for instance) to get

$$\frac{S^{(1)}(t) - S^{(2)}(t)}{S^{(2)}(t)} = -S^{(1)}(t)\Delta_1(t) + \int_0^t \Delta_1(u)S^{(1)}(\mathrm{d}u).$$

Then, as $S^{(2)}(t) \leq 1$, we obtain that

$$\begin{aligned}\left|S^{(1)}(t) - S^{(2)}(t)\right| &\leq \left(S^{(1)}(t)\left|\Delta_1(t)\right| + \left(1 - S^{(1)}(t)\right) \sup_{u\in[0,\tau]} \left|\Delta_1(u)\right|\right) \\ &\leq \sup_{u\in[0,\tau]} \left|\Delta_1(u)\right|. \end{aligned}$$

We conclude by using Lemma 2.18 with $\mathrm{d}\nu = \mathrm{d}\left(\Lambda^{(1)} - \Lambda^{(2)}\right)$ and $h = 1/S^{(2)}$. □

Lemma 2.20. *Let $0 < \theta_1, \theta_2 < 1$ and $\tau > 0$. For $k \in \{1,\ 2\}$, define*

$$\lambda^{(k)}(t) = \int_0^t \frac{G^{(k)}(\mathrm{d}u)}{H^{(k)}(u)},$$

where $G^{(k)} : [0,\tau] \mapsto [0,\beta]$ is càdlàg non-decreasing and $H^{(k)} : [0,\tau] \to [\theta_k, 1]$ is Borelian non-increasing. Then, we have:

$$\left\|\Lambda^{(1)} - \Lambda^{(2)}\right\|_{[0,\tau]} \leq \frac{2}{\theta_1}\left\|G^{(1)} - G^{(2)}\right\|_{[0,\tau]} + \frac{\beta}{\theta_1\theta_2}\left\|H^{(1)} - H^{(2)}\right\|_{[0,\tau]}.$$

Proof. Let $t \in [0,\tau]$. Observe that, by triangular inequality,

$$\begin{aligned}\left|\Lambda^{(1)}(t) - \Lambda^{(2)}(t)\right| &= \left|\int_0^t \frac{\mathrm{d}(G^{(1)} - G^{(2)})}{H^{(1)}} + \int_0^t \frac{H^{(2)} - H^{(1)}}{H^{(1)}H^{(2)}}\,\mathrm{d}G^{(2)}\right| \\ &\leq \frac{2}{\theta_1}\left\|G^{(1)} - G^{(2)}\right\|_{[0,\tau]} + \frac{\beta}{\theta_1\theta_2}\left\|H^{(2)} - H^{(1)}\right\|_{[0,\tau]}, \end{aligned}$$

where the bound for the second term on the right-hand side is straightforward and that for the first term can be deduced from the application of Lemma 2.18 with the measure ν equal to $A \mapsto \int_A \mathrm{d}(G^{(1)} - G^{(2)})$ and the function h equal to $1/H^{(1)}$. □

Lemma 2.21. *Let $\tau > 0$. Let $S^{(1)}$ and $S^{(2)}$ be càdlàg non-increasing functions on $\mathbb{R}_+$ such that $S^{(1)}(0) = S^{(2)}(0) = 1$ and $S^{(2)}(\tau) \geq \theta > 0$. For $k \in \{1, 2\}$, define $\lambda^{(k)}(t) = -\int_0^t S^{(k)}(u-)S^{(k)}(\mathrm{d}u)$ and suppose that*

$$\lambda^{(k)}(t) = \int_0^t \frac{G^{(k)}(\mathrm{d}u)}{H^{(k)}(u)},$$

where $G^{(k)} : [0, \tau] \to [0, \beta]$ and $H^{(k)} : [0, \tau] \to [\theta, 1]$ are respectively non-decreasing and non-increasing borelian functions. Then, there exists a constant $C_{\theta,\beta} > 0$, depending only on θ and β, such that

$$\sup_{t\in[0,\tau]} \left| \int_0^t \frac{S^{(1)}(u-) - S^{(2)}(u-)}{S^{(2)}(u)} \left(\Lambda^{(1)}(\mathrm{d}u) - \Lambda^{(2)}(\mathrm{d}u)\right) \right| \leq C_{\theta,\beta} \left(\left\| H^{(1)} - H^{(2)} \right\|^2_{[0,\tau]} + \left\| G^{(1)} - G^{(2)} \right\|^2_{[0,\tau]} + \|W\|_{[0,\tau]} \right),$$

where

$$W(t) = \int_{u=0}^t \int_{s=0}^u \frac{S^{(2)}(s-)\left(G^{(1)}(\mathrm{d}s) - G^{(2)}(\mathrm{d}s)\right)}{S^{(2)}(s)H^{(2)}(s)} \frac{\mathrm{d}\left(G^{(1)}(\mathrm{d}u) - G^{(2)}(\mathrm{d}u)\right)}{S^{(2)}(u)H^{(2)}(u)}.$$

Proof. The proof consists in showing first that there exist constants $C_{1,\theta,\beta}$ and $C_{2,\theta,\beta}$ such that

$$\sup_{t\in[0,\tau]} \left| \int_0^t \frac{\hat{S}^{(1)}(u-) - S^{(2)}(u-)}{S^{(2)}(u)} \left(\Lambda^{(1)}(\mathrm{d}u) - \Lambda^{(2)}(\mathrm{d}u)\right) \right| \leq C_{1,\theta,\beta} \left(\left\| G^{(1)} - G^{(2)} \right\|^2_{[0,\tau]} + \left\| H^{(1)} - H^{(2)} \right\|^2_{[0,\tau]} \right) + \|\Pi\|_{[0,\tau]}, \quad (2.44)$$

where

$$\begin{aligned} \Pi(t) &= \int_0^t \Delta_2(u)\Delta_1(\mathrm{d}u), \\ \Delta_2(t) &= \int_0^t S^{(2)}(u-)\Delta_1(\mathrm{d}u), \\ \Delta_1(t) &= \int_0^t S^{(2)}(u)^{-1}\Delta(\mathrm{d}u), \\ \Delta &= \Lambda^{(1)} - \Lambda^{(2)}, \end{aligned}$$

and next that

$$|\Pi - W|_{[0,\tau]} \leq C_{2,\theta,\beta} \left(\left| H^{(1)} - H^{(2)} \right|^2_{[0,\tau]} + \left| G^{(1)} - G^{(2)} \right|^2_{[0,\tau]} \right). \tag{2.45}$$

In order to establish eq. (2.44), we successively apply eq. (2.43), Fubini's theorem and the integration by part formula:

$$\begin{aligned}
&\int_{u=0}^{t} \left(S^{(1)}(u-) - S^{(2)}(u-) \right) \Delta_1(\mathrm{d}u) \\
&\quad = -\int_{u=0}^{t} \int_{v=0}^{u-} S^{(1)}(v-)\Delta_1(\mathrm{d}v) S^{(2)}(u-)\Delta_1(\mathrm{d}u) \\
&\quad = -\int_{v=0}^{t} \left(\int_{u=v}^{t} S^{(2)}(u-)\Delta_1(\mathrm{d}u) \right) S^{(1)}(v-)\Delta_1(\mathrm{d}v) \\
&\quad = -\Delta_2(t) \int_0^t S^{(1)}(v-)\Delta_1(\mathrm{d}v) + \int_0^t S^{(1)}(v-)\Pi(\mathrm{d}v) \\
&\quad = -\Delta_2(t) \left(S^{(1)}(t)\Delta_1(t) - \int_0^t \Delta_1(u) S^{(1)}(\mathrm{d}u) \right) \\
&\qquad\qquad + S^{(1)}(t)\Pi(t) - \int_0^t \Pi(u) S^{(1)}(\mathrm{d}u) \\
&\quad \leq 2 \left\| \Delta_2 \right\|_{[0,\tau]} \left\| \Delta_1 \right\|_{[0,\tau]} + 2 \left\| \Pi \right\|_{[0,\tau]}.
\end{aligned} \tag{2.46}$$

From eq. (2.42), we deduce that $\|\Delta_2\|_{[0,\tau]} \leq \|\Delta_1\|_{[0,\tau]}$ (because $S^{(2)}\mathbb{1}_{[0,\tau]}$ belongs to the space DE) and, from Lemma 2.18, it follows that $\|\Delta_1\|_{[0,\tau]} \leq 2\theta^{-1}\|\delta\|_{[0,\tau]}$. Apply next Lemma 2.20 to obtain

$$\left\| \Delta_2 \right\|_{[0,\tau]} \left\| \Delta_1 \right\|_{[0,\tau]} \leq \frac{8}{\theta^2} \left(\frac{4}{\theta^2} \left\| G^{(1)} - G^{(2)} \right\|^2_{[0,\tau]} + \frac{\beta^2}{\theta^4} \left\| H^{(1)} - H^{(2)} \right\|^2_{[0,\tau]} \right).$$

Combined with eq. (2.46), this proves eq. (2.44). For eq. (2.45), the application of the Taylor expansion

$$\frac{1}{x} = \frac{1}{a} - \frac{x-a}{a^2} + \frac{(x-a)^2}{xa^2} \tag{2.47}$$

yields

$$\mathrm{d}\Delta = \frac{\mathrm{d}\left(G^{(1)} - G^{(2)} \right)}{H^{(2)}} - \frac{\left(H^{(1)} - H^{(2)} \right) \mathrm{d}G^{(1)}}{(H^{(2)})^2} + \frac{\left(H^{(1)} - H^{(2)} \right)^2 \mathrm{d}G^{(1)}}{(H^{(2)})^2 H^{(1)}}.$$

Set $c(s) = S^{(2)}(s-)/S^{(2)}(s)$. It follows that

$$\Pi(t) = \int_{u=0}^{t}\int_{s=0}^{u} c(s)\Bigg(\frac{G^{(1)}(\mathrm{d}s) - G^{(2)}(\mathrm{d}s)}{H^{(2)}(s)} - \frac{\left(H^{(1)}(s) - H^{(2)}(s)\right)G^{(1)}(\mathrm{d}s)}{H^{(2)}(s)^2} + \frac{\left(H^{(1)}(s) - H^{(2)}(s)\right)^2 G^{(1)}(\mathrm{d}s)}{H^{(2)}(s)^2 H^{(1)}(s)}\Bigg)\Delta_1(\mathrm{d}u).$$

Observe that

$$\begin{aligned}
\Pi(t) - W(t) = & \\
&- \int_{u=0}^{t}\int_{s=0}^{u} c(s)\frac{G^{(1)}(\mathrm{d}s) - G^{(2)}(\mathrm{d}s)}{H^{(2)}(s)}\frac{\left(H^{(1)}(u) - H^{(2)}(u)\right)G^{(1)}(\mathrm{d}u)}{S^{(2)}(u)H^{(1)}(u)H^{(2)}(u)} \\
&+ \int_{u=0}^{t}\int_{s=0}^{u} c(s)\frac{\left(H^{(1)}(s) - H^{(2)}(s)\right)G^{(1)}(\mathrm{d}s)}{H^{(2)}(s)^2} \\
&\qquad \times \frac{\left(H^{(1)}(u) - H^{(2)}(u)\right)G^{(1)}(\mathrm{d}u)}{S^{(2)}(u)H^{(1)}(u)H^{(2)}(u)} \\
&- \int_{u=0}^{t}\int_{s=0}^{u} c(s)\frac{\left(H^{(1)}(s) - H^{(2)}(s)\right)G^{(1)}(\mathrm{d}s)}{H^{(2)}(s)^2}\frac{G^{(1)}(\mathrm{d}u) - G^{(2)}(\mathrm{d}u)}{S^{(2)}(u)H^{(2)}(u)} \\
&+ \int_{u=0}^{t}\int_{s=0}^{u} \frac{\left(H^{(1)}(s) - H^{(2)}(s)\right)^2 G^{(1)}(\mathrm{d}s)}{H^{(2)}(s)^2 H^{(1)}(s)}\Delta_1(\mathrm{d}u) \\
= & A + B + C + D.
\end{aligned}$$

We next bound each term on the right-hand side of the equation above. Successively apply Lemma 2.18 and eq. (2.42) to get

$$\begin{aligned}
&\left|\int_0^u c(s)\frac{G^{(1)}(\mathrm{d}s) - G^{(2)}(\mathrm{d}s)}{H^{(2)}(s)}\right| \\
&\quad \le \frac{2}{\theta^2}\sup_u\left|\int_0^u S^{(2)}(s-)\left(G^{(1)}(\mathrm{d}s) - G^{(2)}(\mathrm{d}s)\right)\right| \\
&\quad = \frac{2}{\theta^2}\sup_u\left|\int S^{(2)}(s-)\mathbb{1}_{s\le u}\left(G^{(1)}(\mathrm{d}s) - G^{(2)}(\mathrm{d}s)\right)\right| \\
&\quad \le \frac{2}{\theta^2}\left\|G^{(1)} - G^{(2)}\right\|_{[0,\tau]}.
\end{aligned}$$

Because, for any $u \in [0, \tau]$ we have that

$$\frac{1}{S^{(2)}(u)H^{(1)}(u)H^{(2)}(u)} \le \frac{1}{\theta^3},$$

we can write

$$\begin{aligned}
|A| &\leq \frac{1}{\theta^3}\int_{u=0}^{t}\left|\int_{s=0}^{u} c(s)\frac{G^{(1)}(\mathrm{d}s)-G^{(2)}(\mathrm{d}s)}{H^{(2)}(s)}\right| \\
&\qquad\qquad \times\left|H^{(1)}(u)-H^{(2)}(u)\right| G^{(1)}(\mathrm{d}u) \\
&\leq \frac{1}{2\theta^3}\int_{u=0}^{t}\left(\int_{0}^{u} c(s)\frac{G^{(1)}(\mathrm{d}s)-G^{(2)}(\mathrm{d}s)}{H^{(2)}(s)}\right)^2 \\
&\qquad\qquad + \left(H^{(1)}(u)-H^{(2)}(u)\right)^2 G^{(1)}(\mathrm{d}u) \\
&\leq \beta\left(\frac{2}{\theta^7}\left\|G^{(1)}-G^{(2)}\right\|_{[0,\tau]}^2 + \left\|H^{(1)}-H^{(2)}\right\|_{[0,\tau]}^2\right).
\end{aligned}$$

In addition, because for any $u \in [0,\tau]$, $c(u)/(H^{(2)}(u))^2 \leq 1/\theta^3$ we have: $\forall t \in [0,\tau]$,

$$\begin{aligned}
|B| &\leq \left(\frac{1}{\theta^3}\right)^2 \int_{u=0}^{t}\int_{s=0}^{t}\left|H^{(1)}(s)-H^{(2)}(s)\right| G^{(1)}(\mathrm{d}s) \\
&\qquad\qquad \times\left|H^{(1)}(u)-H^{(2)}(u)\right| G^{(1)}(\mathrm{d}u) \\
&= \frac{1}{\theta^6}\left(\int_{s=0}^{t}\left|H^{(1)}(s)-H^{(2)}(s)\right| G^{(1)}(\mathrm{d}s)\right)^2 \\
&\leq \frac{\beta^2}{\theta^6}\left\|H^{(1)}-H^{(2)}\right\|_{[0,\tau]}^2.
\end{aligned}$$

If we define

$$\gamma_2(t) = \int_0^t \frac{G^{(1)}(\mathrm{d}u)-G^{(2)}(\mathrm{d}u)}{S^{(2)}(u)H^{(2)}(u)},$$

we can apply Fubini's theorem and get

$$\begin{aligned}
|C| &= \left|\int_{u=0}^{t}\int_{s=0}^{u} c(s)\frac{\left(H^{(1)}(s)-H^{(2)}(s)\right)G^{(1)}(\mathrm{d}s)}{H^{(2)}(s)^2}\frac{G^{(1)}(\mathrm{d}u)-G^{(2)}(\mathrm{d}u)}{S^{(2)}(u)H^{(2)}(u)}\right| \\
&= \left|\int_{s=0}^{t}\int_{u=s}^{t} \frac{G^{(1)}(\mathrm{d}u)-G^{(2)}(\mathrm{d}u)}{S^{(2)}(u)H^{(2)}(u)}c(s)\frac{\left(H^{(1)}(s)-H^{(2)}(s)\right)G^{(1)}(\mathrm{d}s)}{(H^{(2)}(s))^2}\right| \\
&\leq \frac{1}{\theta^3}\int_{s=0}^{t}\left|\gamma_2(t)-\gamma_2(s)\right| \times\left|H^{(1)}(s)-H^{(2)}(s)\right| G^{(1)}(\mathrm{d}s) \\
&\leq \frac{2\beta}{\theta^3}\left\|\gamma_2\right\|_{[0,\tau]} \times\left\|H^{(1)}-H^{(2)}\right\|_{[0,\tau]}.
\end{aligned}$$

Then, using Lemma 2.18, it follows that

$$\begin{aligned}
|C| &\leq \frac{4\beta}{\theta^6}\left\|G^{(1)}-G^{(2)}\right\|_{[0,\tau]}\left\|H^{(1)}-H^{(2)}\right\|_{[0,\tau]} \\
&\leq \frac{2\beta}{\theta^6}\left(\left\|G^{(1)}-G^{(2)}\right\|_{[0,\tau]}^2 + \left\|H^{(1)}-H^{(2)}\right\|_{[0,\tau]}^2\right).
\end{aligned}$$

The last term can be treated by means of Fubini's theorem. Indeed, because $\|\Delta_1\|_{[0,\tau]} \leq 2\beta/\theta$ and for any $u \in [0,\tau]$, $1/(H^{(2)}(u)^2 H^{(1)}(u)) \leq 1/\theta^3$, we have

$$\begin{aligned}
|D| &= \left| \int_{u=0}^{t} \int_{s=0}^{u} \frac{\left(H^{(1)}(s) - H^{(2)}(s)\right)^2 G^{(1)}(\mathrm{d}s)}{(H^{(2)}(s))^2 H^{(1)}(s)} \Delta_1(\mathrm{d}u) \right| \\
&\leq \int_{s=0}^{t} \left| \left(\int_{u=s}^{t} \Delta_1(\mathrm{d}u) \right) \frac{\left(H^{(1)}(s) - H^{(2)}(s)\right)^2 G^{(1)}(\mathrm{d}s)}{H^{(2)}(s)^2 H^{(1)}(s)} \right| \\
&\leq \frac{2\beta}{\theta^3} \|\Delta_1\|_{[0,\tau]} \times \left\| H^{(1)} - H^{(2)} \right\|_{[0,\tau]}^2 \\
&\leq \frac{4\beta^2}{\theta^4} \left\| H^{(1)} - H^{(2)} \right\|_{[0,\tau]}^2 .
\end{aligned}$$

Putting all this together, the triangular inequality leads to eq. (2.45) . □

2.8.3 Proof of Theorem 2.6

We start by establishing 3 useful lemmas, namely Lemmas 2.22 to 2.24. Then the proof will follow easily. Define

$$\begin{aligned}
H_{0,h}(y,x) &= \mathbb{E}\left[\hat{H}_{0,n}(y,x)\right], \\
H_h(y,x) &= \mathbb{E}\left[\hat{H}_n(y,x)\right],
\end{aligned}$$

and

$$\begin{aligned}
H_0(y,x) &= H_0(y \mid x) g(x), \\
H(y,x) &= H(y \mid x) g(x).
\end{aligned}$$

Lemma 2.22. *Under Assumption 2.2, there exists $C_0 > 0$ depending only on K and L such that for all $h > 0$,*

$$\begin{aligned}
\sup_{(t,x) \in \mathbb{R}_+ \times \mathbb{R}^d} |H_{0,h}(t,x) - H_0(t,x)| &\leq C_0 h^2, \qquad (2.48) \\
\sup_{(t,x) \in \mathbb{R}_+ \times \mathbb{R}^d} |H_h(t,x) - H(t,x)| &\leq C_0 h^2.
\end{aligned}$$

Proof. The proof results from the application of Lemma 2.17 combined with the smoothness assumptions stipulated. □

Lemma 2.23. *Under Assumption 2.2, There exist constants $M_1 > 0$ and $h_0 > 0$ depending only on K and L such that:*

$$\mathbb{P}\left(\sup_{(t,x)\in\mathbb{R}_+\times\mathbb{R}^d}|\hat{H}_{0,n}(t,x)-H_{0,h}(t,x)|\leq\sqrt{\frac{M_1\,|\log(\varepsilon h^{d/2})|}{nh^d}}\right)\geq 1-\varepsilon,$$
$$\mathbb{P}\left(\sup_{(t,x)\in\mathbb{R}_+\times\mathbb{R}^d}|\hat{H}_{n}(t,x)-H_{h}(t,x)|\leq\sqrt{\frac{M_1\,|\log(\varepsilon h^{d/2})|}{nh^d}}\right)\geq 1-\varepsilon,$$

provided that $h\leq h_0$ and $M_1|\log(\varepsilon h^{d/2})|\leq nh^d$.

Proof. The exponential inequalities stated above directly result from the application of Corollary 2.12 to the uniformly bounded VC-type classes (see Lemmas 2.13 and 2.16)

$$\left\{(y,x')\mapsto K\left(\frac{x-x'}{h}\right)\mathbb{1}_{y>u}:(x,u,h)\in\mathbb{R}^d\times\mathbb{R}_+\times\mathbb{R}_+^*\right\},$$
$$\left\{(y,\delta,x')\mapsto K\left(\frac{x-x'}{h}\right)\mathbb{1}_{y>u,\delta=0}:(x,u,h)\in\mathbb{R}^d\times\mathbb{R}_+\times\mathbb{R}_+^*\right\},$$

whose VC constants are independent from h, with constant envelope $||K||_\infty$, with $k=1$ and $\sigma^2=c_{K,L}^2h^d$ with $c_{K,L}=\sqrt{L\int K^2(x)\,\mathrm{d}x}$. This gives that

$$\mathbb{P}\left(\sup_{(t,x)\in\mathbb{R}_+\times\mathbb{R}^d}|\hat{H}_{0,n}(t,x)-H_{0,h}(t,x)|\leq t\right)\geq 1-\varepsilon,$$
$$\mathbb{P}\left(\sup_{(t,x)\in\mathbb{R}_+\times\mathbb{R}^d}|\hat{H}_{n}(t,x)-H_{h}(t,x)|\leq t\right)\geq 1-\varepsilon,$$

with

$$t=\frac{c_{K,L}}{\sqrt{nh^d}}\left(\left(\frac{1}{C_3}\log\left(\frac{C_2}{\varepsilon}\right)\right)^{1/2}+C_1\left(\log\left(\frac{2\|K\|_\infty}{c_{K,L}h^{d/2}}\right)\right)^{1/2}\right),$$

provided that $h^{d/2}c_{K,L}\leq||K||_\infty$ and

$$\frac{\|K\|_\infty^2}{c_{K,L}^2}\left(\frac{1}{C_3}\log\left(\frac{C_2}{\varepsilon}\right)+C_1^2\log\left(\frac{2\|K\|_\infty}{c_{K,L}h^{d/2}}\right)\right)\leq nh^d.$$

Since, for any positive numbers a,b,γ, it holds that $a^\gamma+b^\gamma\leq 2^\gamma(a+b)^\gamma$, we find, taking h_0 sufficiently small, that

$$t^2\leq\frac{\widetilde{M}_1\left|\log\left(\varepsilon h^{d/2}\right)\right|}{nh^d},$$

for some constant $\widetilde{M}_1 > 0$. Finally, taking h_0 sufficiently small ensures that

$$\frac{\log(C_2)}{C_3} + C_1^2 \log\left(2\frac{|K|_\infty}{c_{K,L}}\right) \leq C_1^2 \log\left(\frac{1}{h^{d/2}}\right),$$

for any $h \leq h_0$, which permits to ensure that the previous condition is satisfied whenever $\widetilde{M}_2|\log(\varepsilon h^{d/2})| \leq nh^d$, for some $\widetilde{M}_2 > 0$. Take $M_1 = \widetilde{M}_1 + \widetilde{M}_2$ to obtain the desired result. □

Lemma 2.24. *Suppose that Assumptions 2.1 and 2.2 are fulfilled. There exist constants $M_1 > 0$ and $h_0 > 0$ depending only on b, L and K such that:*

$$\mathbb{P}\left(\inf_{(t,x)\in\gamma_b} \hat{H}_n(t,x) \geq \frac{3}{4}b^3\right) \geq 1 - \varepsilon,$$

provided that $h \leq h_0$ and $M_1|\log(\varepsilon h^{d/2})| \leq nh^d$.

Proof. Define

$$\mathcal{A}_n = \left\{\sup_{(t,x)\in\gamma_b} |H(t,x) - \hat{H}_n(t,x)| \leq b^3/4\right\}.$$

By virtue of Assumption 2.1, for any $(t,x) \in \gamma_b$, we have:

$$\begin{aligned} H(t \mid x) &= S_C(t \mid x) S_Y(t \mid x) \\ &\geq b^2, \end{aligned}$$

as a consequence of

$$\begin{aligned} \hat{H}_n(t,x) &\geq H(t,x) - |H(t,x) - \hat{H}_n(t,x)|, \\ \mathcal{A}_n &\subset \left\{\inf_{(t,x)\in\gamma_b} \hat{H}_n(t,x) \geq \frac{3}{4}b^3\right\}. \end{aligned}$$

Hence we only have to prove that event $\mathcal{A}_n$ occurs with probability $1 - \varepsilon$ at least. By virtue of Lemma 2.22, as soon as $h \leq \sqrt{3b^3/(8C_0)}$, we have

$$\sup_{(t,x)\in\gamma_b} |H_h(t,x) - H(t,x)| \leq \frac{3}{8}b^3,$$

and thus

$$\left\{\sup_{(t,x)\in\gamma_b} |\hat{H}_n(t,x) - H_h(t,x)| \leq \frac{3}{8}b^3\right\} \subset \mathcal{A}_n.$$

Simply use Lemma 2.23 to ensure that the event in the left-hand side holds with probability $1 - \varepsilon$ whenever $M_1|\log(\varepsilon h^{d/2})| \leq nh^d$ (where M_1 now depends on b, L and K) and $h \leq h_0$. □

Now we conclude the proof. We start by using Lemma 2.24 to get that $\inf_{(t,x)\in\gamma_b} \hat{H}_n(t,x) \geq 3b^3/4$ happens with probability $1-\varepsilon/3$. We suppose that this event is realized in the following. Let $(t,x) \in \gamma_b$ and define

$$\tau_x = \inf\{t \geq 0 : \min\{S_C(t \mid x), S_Y(t \mid x)\} > b\}.$$

Observing that the choice of kernel K guarantees that $\hat{S}_{C,n}(\cdot|x)$ is a (random) survival function, we first apply Lemma 2.19 with $S^{(1)} = \hat{S}_{C,n}(\cdot|x)$, $S^{(2)} = S_C(\cdot|x)$ and $\theta = b$ to get:

$$\left\|\hat{S}_{C,n}(\cdot|x) - S_C(\cdot|x)\right\|_{[0,\tau_x]} \leq \frac{2}{b}\left\|\hat{\lambda}_{C,n}(\cdot|x) - \lambda_C(\cdot|x)\right\|_{[0,\tau_x]}. \tag{2.49}$$

Applying Lemma 2.20 with

$$\begin{aligned}
\Lambda^{(1)}(u) &= \lambda_C(u \mid x) \\
&= -\int_0^u \frac{H_0(\mathrm{d}s, x)}{H(s-, x)}, \\
\Lambda^{(2)}(u) &= \hat{\lambda}_{C,n}(u \mid x) \\
&= -\int_0^u \frac{\hat{H}_{0,n}(\mathrm{d}s, x)}{\hat{H}_n(s-, x)}, \\
\beta &= 1, \\
\theta_1 &= b^3 \leq H(s,x), \\
\theta_2 &= \frac{3}{4}b^3,
\end{aligned}$$

because $\inf_{(t,x)\in\gamma_b} \hat{H}_n(t,x) \geq b^3/4$, next yields

$$\begin{aligned}
&\left\|\hat{\lambda}_{C,n}(\cdot \mid x) - \lambda_C(\cdot \mid x)\right\|_{[0,\tau_x]} \\
&\quad\leq \frac{2}{b^3}\left\|\hat{H}_{0,n}(\cdot, x) - H_0(\cdot \mid x)g(x)\right\|_{[0,\tau_x]} \\
&\qquad + \frac{4}{3b^6}\left\|\hat{H}_n(\cdot, x) - H(\cdot \mid x)g(x)\right\|_{[0,\tau_x]}.
\end{aligned} \tag{2.50}$$

Combining eq. (2.49) and eq. (2.50), using Lemma 2.22 and taking the supremum over x such that $g(x) > b$, we obtain that, the following bound

holds true:

$$\begin{aligned}
&\sup_{(t,x)\in\gamma_b}\left|\hat{S}_{C,n}(t\mid x)-S_C(t\mid x)\right| \\
&\quad\le \frac{4}{b^4}\sup_{(t,x)\in\gamma_b}\left|\hat{H}_{0,n}(t,x)-H_0(t,x)\right| \\
&\qquad+\frac{8}{3b^7}\sup_{(t,x)\in\gamma_b}\left|\hat{H}_n(t,x)-H(t,x)\right| \\
&\quad\le \frac{4}{b^4}\sup_{(t,x)\in\gamma_b}\left|\hat{H}_{0,n}(t,x)-H_{0,h}(t,x)\right|+\frac{4}{b^4}C_0h^2 \\
&\qquad+\frac{8}{3b^7}\sup_{(t,x)\in\gamma_b}\left|\hat{H}_n(t,x)-H_h(t,x)\right|+\frac{8}{3b^7}C_0h^2.
\end{aligned} \tag{2.51}$$

Lemma 2.23 with the probability level $\varepsilon/3$ allows us to bound the 2 previous random terms. Combined with the union bound (with 3 events having probability smaller than $\varepsilon/3$), permits claiming that with probability greater than $1-\varepsilon$:

$$\begin{aligned}
&\sup_{(t,x)\in\gamma_b}\left|\hat{S}_{C,n}(t\mid x)-S_C(t\mid x)\right| \\
&\qquad\le \frac{4}{b^4}\left(1+\frac{2}{3b^3}\right)\left(C_0h^2+\sqrt{\frac{M_1\left|\log(\varepsilon h^{d/2})\right|}{nh^d}}\right),
\end{aligned}$$

provided that (to apply Lemma 2.23) $h\le h_0$ and $nh^d\ge M_1|\log(3\varepsilon h^{d/2})|$. Examining the different terms and taking h_0 small enough lead to the stated result.

2.8.4 Proof of Proposition 2.7

Proof of (i): Observe that: $\forall i\in\{1,\ldots,n\}$,

$$\sup_{(t,x)\in\mathbb{K}}|\hat{H}^{(i)}_{0,n}(t,x)-\hat{H}_{0,n}(t,x)|\le 2\frac{\|K\|_\infty}{(n-1)h^d}, \tag{2.52}$$

$$\sup_{(t,x)\in\mathbb{K}}|\hat{H}^{(i)}_{n}(t,x)-\hat{H}_{n}(t,x)|\le 2\frac{\|K\|_\infty}{(n-1)h^d}. \tag{2.53}$$

The result follows from the union bound and that each of these events

$$\mathcal{B}_n^{(1)}\overset{\text{def}}{=}\bigcap_{i\le n}\left\{\forall(t,x)\in\mathbb{K},\ \hat{H}^{(i)}_n(t,x)\ge\frac{b^3}{2}\right\},$$

$$\mathcal{B}_n^{(2)}\overset{\text{def}}{=}\bigcap_{i\le n}\left\{\forall(t,x)\in\mathbb{K},\ \hat{S}^{(i)}_{C,n}(t,x)\ge\frac{b}{2}\right\},$$

has probability $1-\varepsilon/2$ under the mentioned condition on (n,h). Apply Lemma 2.24 to choose (n,h) such that with probability $1-\varepsilon/2$,

$$\inf_{(t,x)\in\mathbb{K}} \hat{H}_n(t,x) \geq \frac{3}{4}b^3.$$

Using eq. (2.53) and the triangle inequality, we get that $\mathcal{B}_n^{(1)}$ has probability $1-\varepsilon/2$ provided that

$$2\frac{\|K\|_\infty}{(n-1)h^d} \leq \frac{b^3}{4}.$$

Suppose that event $\mathcal{B}_n^{(1)}$ is realized. The same reasoning as that used in the proof of Theorem 2.6 (see eqs. (2.49) to (2.51)), with

$$\begin{aligned} S^{(1)}(\cdot) &= S_C(\cdot|x), \\ S^{(2)}(\cdot) &= S_{C,n}^{(i)}(\cdot|x), \\ \beta &= 1, \\ \theta_1 &= b^3, \\ \theta_2 &= b^3/2, \end{aligned}$$

(as $\mathcal{B}_n^{(1)}$ is realized), combined with the triangular inequality, yields: $\forall i \in \{1,\ldots,n\}$,

$$\begin{aligned} &\sup_{(t,x)\in\mathbb{K}} \left|\hat{S}_{C,n}^{(i)}(t|x) - S_C(t|x)\right| \\ &\leq \frac{4}{b^4}\left(\sup_{(t,x)\in\mathbb{K}} \left|\hat{H}_{0,n}^{(i)}(t,x) - \hat{H}_{0,n}(t,x)\right| + \sup_{(t,x)\in\mathbb{K}} \left|\hat{H}_{0,n}(t,x) - H_0(t,x)\right|\right) \\ &\quad + \frac{4}{b^7}\left(\sup_{(t,x)\in\mathbb{K}} \left|\hat{H}_n^{(i)}(t,x) - \hat{H}_n(t,x)\right| + \sup_{(t,x)\in\mathbb{K}} \left|\hat{H}_n(t,x) - H(t,x)\right|\right). \end{aligned}$$

We further assume that

$$2\left(\frac{4}{b^4}+\frac{4}{b^7}\right)\frac{\|K\|_\infty}{(n-1)h^d} \leq \frac{b}{4},$$

which is realized whenever h_0 is small enough and M_1, appearing in the condition $nh^d \geq M_1|\log(h^{d/2}\varepsilon)|$, is large enough. From $\mathbb{K} \subset \gamma_b$ and eq. (2.52)-eq. (2.53), it results that

$$\left\{\sup_{(t,x)\in\mathbb{K}} |\hat{H}_{0,n}(t,x) - H_0(t,x)| \leq \frac{b^5}{32}\right\} \bigcap \left\{\sup_{(t,x)\in\mathbb{K}} |\hat{H}_n(t,x) - H(t,x)| \leq \frac{b^8}{32}\right\},$$

is included in the set $\mathcal{B}_n^{(2)}$. Following the treatment of eq. (2.51), it is easy to see that the latter event occurs with probability $1-\varepsilon/2$ whenever $h \geq h_0$ is small enough (for the bias) and $nh^d \geq M_1|\log(h^{d/2}\varepsilon)|$.

Proof of (ii). We suppose that the event $\mathcal{E}_n$ is realized. For all $i \in \{1, \ldots, n\}$, recall that

$$\hat{\Lambda}_{C,n}^{(i)}(\mathrm{d}u \mid x) = -\frac{\hat{H}_{0,n}^{(i)}(\mathrm{d}u, x)}{\hat{H}_n^{(i)}(u-, x)},$$
$$\hat{\Delta}_n^{(i)} = \hat{\Lambda}_{C,n}^{(i)} - \Lambda_C,$$

and that $c(s \mid x) = S_C(s- \mid x)/S_C(s \mid x)$. It results from Theorem 3.2.3 in Fleming and Harrington (1991), page 97 that

$$\frac{\hat{S}_{C,n}^{(i)}(t \mid x) - S_C(t \mid x)}{S_C(t \mid x)} =$$
$$-\int_0^t c(u \mid x)\hat{\Delta}_n^{(i)}(\mathrm{d}u|x) - \int_0^t \frac{\hat{S}_{C,n}^{(i)}(u- \mid x) - S_C(u- \mid x)}{S_C(u \mid x)}\hat{\Delta}_n^{(i)}(\mathrm{d}u|x).$$

The Taylor expansion of eq. (2.47) gives that

$$\hat{\Delta}_n^{(i)}(\mathrm{d}u \mid x) = \frac{\hat{H}_{0,n}^{(i)}(\mathrm{d}u, x) - H_0(\mathrm{d}u, x)}{H(u, x)}$$
$$- \frac{\left(\hat{H}_n^{(i)}(u, x) - H(u, x)\right)\hat{H}_{0,n}^{(i)}(\mathrm{d}u, x)}{H(u, x)^2}$$
$$+ \frac{\left(\hat{H}_n^{(i)}(u, x) - H(u, x)\right)^2 \hat{H}_{0,n}^{(i)}(\mathrm{d}u, x)}{H(u, x)^2\hat{H}_n^{(i)}(u, x)},$$

which implies that

$$\frac{\hat{S}_{C,n}^{(i)}(t \mid x) - S_C(t \mid x)}{S_C(t \mid x)} = \hat{a}_n^{(i)}(t \mid x) + \hat{b}_n^{(i)}(t \mid x), \tag{2.54}$$

where

$$\hat{a}_n^{(i)}(t \mid x) = -\int_0^t \frac{c(u \mid x)}{H(u, x)}\left(\hat{H}_{0,n}^{(i)}(\mathrm{d}u, x) - H_0(\mathrm{d}u, x)\right)$$
$$+ \int_0^t \frac{c(u \mid x)}{H(u, x)^2}\left(\hat{H}_n^{(i)}(u, x) - H(u, x)\right)\hat{H}_{0,n}^{(i)}(\mathrm{d}u, x),$$
$$\hat{b}_n^{(i)}(t \mid x) = -\int_0^t \frac{c(u \mid x)}{H(u, x)^2\hat{H}_n^{(i)}(u, x)}\left(\hat{H}_n^{(i)}(u, x) - H(u, x)\right)^2 \hat{H}_{0,n}^{(i)}(\mathrm{d}u, x)$$
$$- \int_0^t \frac{\hat{S}_{C,n}^{(i)}(u- \mid x) - S_C(u- \mid x)}{S_C(u \mid x)}\hat{\Delta}_n^{(i)}(\mathrm{d}u \mid x).$$

Now, using eq. (2.47), we obtain that: $\forall \varphi \in \Phi$,

$$\begin{aligned} Z_n(\varphi) &= \frac{1}{n}\sum_{i=1}^{n}\left(\delta_i \frac{\varphi(T_i, X_i)}{\hat{S}^{(i)}_{C,n}(T_i \mid X_i)} - \mathbb{E}\left[\delta \frac{\varphi(T, X)}{S_C(T \mid X)}\right]\right) \\ &= \frac{1}{n}\sum_{i=1}^{n}\left(\delta_i \frac{\varphi(T_i, X_i)}{S_C(T_i \mid X_i)} - \mathbb{E}\left[\delta \frac{\varphi(T_i, X_i)}{S_C(T \mid X)}\right]\right) \\ &- \frac{1}{n}\sum_{i=1}^{n}\delta_i \varphi(T_i, X_i)\frac{\hat{S}^{(i)}_{C,n}(T_i \mid X_i) - S_C(T_i \mid X_i)}{S^2_C(T_i \mid X_i)} \\ &+ \frac{1}{n}\sum_{i=1}^{n}\delta_i \varphi(T_i, X_i)\frac{\left(S_C(T_i \mid X_i) - \hat{S}^{(i)}_{C,n}(T_i \mid X_i)\right)^2}{S^2_C(T_i \mid X_i)\hat{S}^{(i)}_{C,n}(T_i \mid X_i)}. \end{aligned}$$

Then, using eq. (2.54), we retrieve the expected terms

$$Z_n(\varphi) = L_n(\varphi) + M_n(\varphi) + R_n(\varphi),$$

which proves (*ii*).

2.8.5 Proof of Proposition 2.8

The proof is based on the decomposition stated in Proposition 2.7, combined with the lemmas below that permit to control each term involved in it. Their proofs are given in the next section of the Appendix.

The term $L_n(\varphi)$ is a basic i.i.d. centered average. As shown in the lemma stated below, its uniform fluctuations can be controlled by standard results in empirical process theory.

Lemma 2.25. *Suppose that the hypotheses of Proposition 2.8 are fulfilled. Then, for any $\varepsilon \in]0, 1[$, we have with probability at least $1 - \varepsilon$:*

$$\sup_{\varphi \in \Phi} |L_n(\varphi)| \leq \sqrt{\frac{M_1 \log(M_2/\varepsilon)}{n}},$$

provided that $n \geq M_1 \log(M_2/\varepsilon)$, where $M_1 > 0$ and $M_2 > 1$ are constants depending on (A, v), K, M_Φ, and b only.

We now turn to the term $M_n(\varphi)$. Observe it can be decomposed as

$$M_n(\varphi) = V_{n,1}(\varphi) + B_{n,1}(\varphi) + V_{n,2}(\varphi) + B_{n,2}(\varphi),$$

where

$$V_{n,1}(\varphi) = \frac{1}{n}\sum_{i=1}^{n} \frac{\delta_i \varphi(T_i, X_i)}{S_C(T_i \mid X_i)} \int_0^{T_i} \frac{c(u \mid X_i)}{H(u, X_i)^2} \left(\hat{H}_n^{(i)}(u, X_i) - H_h(u, X_i)\right) \hat{H}_{0,n}^{(i)}(\mathrm{d}u, X_i),$$

$$B_{n,1}(\varphi) = \frac{1}{n}\sum_{i=1}^{n} \frac{\delta_i \varphi(T_i, X_i)}{S_C(T_i \mid X_i)} \int_0^{T_i} \frac{c(u \mid X_i)}{H(u, X_i)^2} \left(H_h(u, X_i) - H(u, X_i)\right) \hat{H}_{0,n}^{(i)}(\mathrm{d}u, X_i),$$

$$V_{n,2}(\varphi) = -\frac{1}{n}\sum_{i=1}^{n} \frac{\delta_i \varphi(T_i, X_i)}{S_C(T_i \mid X_i)} \int_0^{T_i} \frac{c(u \mid X_i)}{H(u, X_i)} \left(\hat{H}_{0,n}^{(i)}(\mathrm{d}u, X_i) - H_{0,h}(\mathrm{d}u, X_i)\right),$$

$$B_{n,2}(\varphi) = -\frac{1}{n}\sum_{i=1}^{n} \frac{\delta_i \varphi(T_i, X_i)}{S_C(T_i \mid X_i)} \int_0^{T_i} \frac{c(u \mid X_i)}{H(u, X_i)} \left(H_{0,h}(\mathrm{d}u, X_i) - H_0(\mathrm{d}u, X_i)\right).$$

Next we treat the bias terms $B_{n,1}$ and $B_{n,2}$.

Lemma 2.26. *Under the assumptions of Proposition 2.8, for any $\varepsilon \in]0,1[$, we have, with probability $1-\varepsilon$:*

$$\sup_{\varphi\in\Phi} |B_{n,1}(\varphi)| \le M_1 h^2,$$
$$\sup_{\varphi\in\Phi} |B_{n,2}(\varphi)| \le M_1 h^2,$$

provided that $n \ge M_2 |\log(h^{d/2}\varepsilon)|$, where $M_1 > 0$, $M_2 > 0$ depend only on M_Φ, K, L and b.

Now we consider $V_{n,1}(\varphi)$. For simplicity, we set $K_{ij} = K_h(X_i - X_j)$ for $1 \le i, j \le n$. We have:

$$\begin{aligned} V_{n,1}(\varphi) &= \frac{1}{n(n-1)} \sum_{\substack{(i,j)\\ i\neq j}} \frac{\delta_i \varphi(T_i, X_i)}{S_C(T_i \mid X_i)} \\ &\qquad \times \mathbb{1}_{T_j \le T_i} \frac{(1-\delta_j) K_{ij} c(T_j \mid X_i)}{H(T_j, X_i)^2} \left(\hat{H}_n^{(i)}(T_j, X_i) - H_h(T_j, X_i)\right) \\ &= \frac{1}{n(n-1)^2} \sum_{\substack{(i,j,k)\\ i\neq j, i\neq k}} v_{i,j,k}(\varphi) \\ &= V'_{n,1}(\varphi) + V''_{n,1}(\varphi), \end{aligned}$$

where, for all $1 \le i, j, k \le n$,

$$v_{i,j,k}(\varphi) = \frac{\delta_i \varphi(T_i, X_i)}{S_C(T_i \mid X_i)} \mathbb{1}_{T_j \le T_i} \frac{(1-\delta_j) K_{ij} c(T_j \mid X_i)}{H(T_j, X_i)^2} \left(\mathbb{1}_{T_k > T_j} K_{ik} - H_h(T_j, X_i)\right),$$

and

$$V'_{n,1}(\varphi) = \frac{1}{n(n-1)^2} \sum_{\substack{(i,j,k) \\ i\neq j, i\neq k, j\neq k}} v_{i,j,k}(\varphi),$$
$$V''_{n,1}(\varphi) = \frac{1}{n(n-1)^2} \sum_{\substack{(i,j) \\ i\neq j}} v_{i,j,j}(\varphi).$$

The lemma stated below provides a uniform bound for $V''_{n,1}(\varphi)$.

Lemma 2.27. *Under the assumptions of Proposition 2.8, we have, with probability* 1*:*

$$\sup_{\varphi\in\Phi} |V''_{n,1}(\varphi)| \leq \frac{M_1}{nh^d},$$

where $M_1 > 0$ *depends only on* M_Φ*,* K*,* L *and* b*.*

We now consider $V'_{n,1}(\varphi)$. Set $Z_k = (X_k, T_k, \delta_k)$ for $k \in \{1, \ldots, n\}$. It can be decomposed as follows:

$$V'_{n,1}(\varphi) = \frac{n-2}{n-1} \left(U^{(1)}_{n,1}(\varphi) + U^{(2)}_{n,1}(\varphi) + U^{(3)}_{n,1}(\varphi) + L'_n(\varphi) \right),$$

with

$$U^{(1)}_{n,1}(\varphi) = \frac{1}{n(n-1)(n-2)} \sum_{\substack{(i,j,k) \\ i\neq j, i\neq k, j\neq k}} \Big(v_{i,j,k}(\varphi) - \mathbb{E}\left[v_{i,j,k}(\varphi) \mid Z_j, Z_k \right] - \mathbb{E}\left[v_{i,j,k}(\varphi) \mid Z_i, Z_k \right] + \mathbb{E}\left[v_{i,j,k}(\varphi) \mid Z_k \right] \Big),$$
$$U^{(2)}_{n,1}(\varphi) = \frac{1}{n(n-1)} \sum_{\substack{(j,k) \\ j\neq k}} \Big(\mathbb{E}\left[v_{i,j,k}(\varphi) \mid Z_j, Z_k \right] - \mathbb{E}\left[v_{i,j,k}(\varphi) \mid Z_k \right] \Big),$$
$$U^{(3)}_{n,1}(\varphi) = \frac{1}{n(n-1)} \sum_{\substack{(i,k) \\ i\neq k}} \Big(\mathbb{E}\left[v_{i,j,k}(\varphi) \mid Z_i, Z_k \right] - \mathbb{E}\left[v_{i,j,k}(\varphi) \mid Z_k \right] \Big),$$
$$L'_n(\varphi) = \frac{1}{n} \sum_k \mathbb{E}\left[v_{i,j,k}(\varphi) \mid Z_k \right],$$

where i, j and k always denote pairwise distinct indexes, with the varying amount of indexes in the summations being the result of the successive marginalizations necessary to obtain degenerate U-processes. Observe that, for all $\varphi \in \Phi$ and pairwise distinct indexes i, j and k in $\{1, \ldots, n\}$, we have with probability one:

$$\mathbb{E}[v_{i,j,k}(\varphi) \mid Z_i, Z_j] = \mathbb{E}[v_{i,j,k}(\varphi) \mid Z_i] = \mathbb{E}[v_{i,j,k}(\varphi) \mid Z_j] = 0.$$

The quantities $U_{n,1}^{(k)}(\varphi)$, $k \in \{1,2,3\}$ are thus degenerate U-statistics of degree 3, 2 and 2 respectively, whereas $L_n'(\varphi)$ is a basic (centred) i.i.d. average. The following result is essentially proved by applying Corollary 2.12, once the complexity assumptions related to the classes of kernels involved in the definition of these degenerate U-processes have been established. It shows that the terms $U_{n,1}^{(k)}(\varphi)$'s are uniformly negligible.

Lemma 2.28. *Suppose that the hypotheses of Proposition 2.8 are fulfilled. There exist constants M_1, M_2 and h_0 depending on (A,v), M_Φ, L, K and b only, such that for any $\varepsilon \in (0,1)$, each of the following events holds true with probability at least $1-\varepsilon$:*

$$\sup_{\varphi\in\Phi}\left|U_{n,1}^{(1)}(\varphi)\right| \leq \left(\frac{M_1\left|\log(\varepsilon h^{d/2})\right|}{nh^d}\right)^{3/2}, \tag{2.55}$$

$$\sup_{\varphi\in\Phi}\left|U_{n,1}^{(2)}(\varphi)\right| \leq \frac{M_1\left|\log(\varepsilon h^{d/2})\right|}{nh^d}, \tag{2.56}$$

$$\sup_{\varphi\in\Phi}\left|U_{n,1}^{(3)}(\varphi)\right| \leq \frac{M_1\left|\log(\varepsilon h^{d/2})\right|}{nh^d}, \tag{2.57}$$

as soon as $h \leq h_0$ and $M_2\left|\log(\varepsilon h^d)\right| \leq nh^{2d}$.

Maximal deviation inequalities for the $L_n'(\varphi)$ can be obtained by means of classical results in empirical process theory, like for $L_n(\varphi)$.

Lemma 2.29. *Suppose that the hypotheses of Proposition 2.8 are fulfilled. Then, for any $\varepsilon \in]0,1[$, we have with probability at least $1-\varepsilon$:*

$$\sup_{\varphi\in\Phi}|L_n'(\varphi)| \leq \sqrt{\frac{M_1\log(M_2/\varepsilon)}{n}},$$

as soon as $M_2\left|\log(\varepsilon h^d)\right| \leq nh^{2d}$ and $h \leq h_0$ where h_0, $M_1 > 0$ and $M_2 > 1$ are constants depending on (A,v), K, M_Φ, L and b only.

The two preceding lemmas combined with the union bound directly yield the following result.

Corollary 2.30. *Suppose that the hypotheses of Proposition 2.8 are fulfilled. There exist constants M_1, M_2, M_3 and h_0 depending on (A,v), M_Φ, L, K and b only such that for any $\varepsilon \in]0,1[$, we have with probability greater than $1-\varepsilon$:*

$$\sup_{\varphi\in\Phi}|V_{n,1}'(\varphi)| \leq M_1\left(\sqrt{\frac{\log(M_2/\varepsilon)}{n}} + \frac{\left|\log(\varepsilon h^{d/2})\right|}{nh^d} + \left(\frac{\left|\log(\varepsilon h^{d/2})\right|}{nh^d}\right)^{3/2}\right),$$

as soon as $h \leq h_0$, $M_3|\log(\varepsilon h^d)| \leq nh^{2d}$.

We next deal with the term $V_{n,2}(\varphi)$.

Lemma 2.31. *Suppose that the hypotheses of Proposition 2.8 are fulfilled. There exist constants M_1, M_2, M_3 and h_0 depending on (A, v), M_Φ, L, K and b only such that for any $\varepsilon \in]0,1[$, we have with probability greater than $1-\varepsilon$:*

$$\sup_{\varphi\in\Phi}|V_{n,2}(\varphi)| \le M_1\left(\sqrt{\frac{\log(M_2/\varepsilon)}{n}} + \frac{\left|\log(\varepsilon h^{d/2})\right|}{nh^{d/2}}\right),$$

as soon as $h \le h_0$, $M_3|\log(\varepsilon h^{d/2})| \le nh^d$.

Finally, we consider the residual $R_n(\varphi)$. Recall first that, for all $\varphi \in \Phi$, we have $R_n(\varphi) = R'_n(\varphi) + R''_n(\varphi)$, where

$$R'_n(\varphi) = -\frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i\varphi(T_i,X_i)}{S_C(T_i \mid X_i)}\hat{b}_n^{(i)}(T_i \mid X_i), \tag{2.58}$$

$$R''_n(\varphi) = \frac{1}{n}\sum_{i=1}^{n}\delta_i\varphi(T_i,X_i)\frac{\left(S_C(T_i \mid X_i) - \hat{S}_{C,n}^{(i)}(T_i \mid X_i)\right)^2}{S_C^2(T_i \mid X_i)\hat{S}_{C,n}^{(i)}(T_i \mid X_i)}. \tag{2.59}$$

Each of the quantities, $R'_n(\varphi)$ and $R''_n(\varphi)$, is treated separately. We start with $R''_n(\varphi)$.

Lemma 2.32. *Suppose that the assumptions of Proposition 2.8 are satisfied. Then, for all $\varepsilon \in]0,1[$, we have with probability greater than $1-\varepsilon$*

$$\sup_{\varphi\in\Phi}|R''_n(\varphi)| \le M_1\left(\frac{\left|\log(\varepsilon h^{d/2})\right|}{nh^d} + \frac{1}{(nh^d)^2} + h^4\right),$$

as soon as $h \le h_0$ and $M_2|\log(\varepsilon h^{d/2})| \le nh^d$, where M_1 and M_2 are nonnegative constants depending on K, L, M_Φ and b only.

We now state a uniform bound for $R'_n(\varphi)$.

Lemma 2.33. *Suppose that the assumptions of Proposition 2.8 are satisfied. Then, for all $\varepsilon \in]0,1[$, we have with probability greater than $1-\varepsilon$*

$$\sup_{\varphi\in\Phi}|R'_n(\varphi)| \le M_1\left(\frac{\left|\log(\varepsilon h^{d/2})\right|}{nh^d} + \frac{\sqrt{\left|\log(\varepsilon h^{d/2})\right|}}{(nh^d)^{3/2}} + \frac{1}{nh^d} + \frac{1}{(nh^d)^2} + h^2\right),$$

as soon as $h \le h_0$ and $M_2|\log(\varepsilon h^{d/2})| \le nh^d$, where M_1 and M_2 are nonnegative constants depending on K, L, M_Φ and b only.

Now we can conclude the proof of Proposition 2.8 by gathering each of the previous results. First note that they all are valid under the condition that

$$\begin{aligned} h &\leq h_0, \\ M_1 \left|\log(\varepsilon h^{d/2})\right| &\leq nh^d, \\ n &\geq M_2 \log\left(\frac{M_3}{\varepsilon}\right). \end{aligned}$$

By taking h_0 small enough, the last requirement is no longer necessary. In addition, if

$$\begin{aligned} nh^d &> 1, \\ \left|\log(\varepsilon h^{d/2})\right| &> 1, \end{aligned}$$

we guarantee that

$$\begin{aligned} \frac{\left|\log(\varepsilon h^{d/2})\right|}{nh^d} &\geq \frac{1}{nh^d} \geq \frac{1}{(nh^d)^2}, \\ \left|\log(\varepsilon h^{d/2})\right|^{1/2} &\leq \left|\log(\varepsilon h^{d/2})\right|^{3/2}, \end{aligned}$$

leading to

$$\frac{\sqrt{\left|\log(\varepsilon h^{d/2})\right|}}{(nh^d)^{3/2}} \leq \left(\frac{\left|\log(\varepsilon h^{d/2})\right|}{nh^d}\right)^{3/2} \leq \frac{\left|\log(\varepsilon h^{d/2})\right|}{nh^d}.$$

Using this manipulation, we obtain the stated result.

2.8.6 Intermediary Results

Here we prove lemmas involved in the argument of the proof of Proposition 2.8. Recall that, under the assumptions stipulated: $\forall (t, x) \in \mathbb{K}$,

$$\begin{aligned} H(t, x) &\geq b^3, \\ S_C(t \mid x) &\geq b, \\ c(t \mid x) &\leq b^{-1}, \\ H_h(t \mid x) &\leq L. \end{aligned} \tag{2.60}$$

Proof of Lemma 2.25 The proof is a direct application of Corollary 2.12 to the i.i.d. sequence $\{(X_n, T_n, \delta_n) : n \geq 1\}$ and the class of functions

$$(x, u, \delta) \in \mathbb{K} \times \{0, 1\} \mapsto \frac{\delta \varphi(u, x)}{S_C(u \mid x)},$$

indexed by $(\varphi, h) \in \Phi\times]0, h_0]$. The previous class is of VC type in virtue of Lemma 2.16. We choose $\sigma = \|G\|_\infty = 2M_\Phi/b$, the bound obtained for $L_n(\varphi)$ is simply

$$\frac{2M_\Phi}{bn^{1/2}}\left((C_1^2\log(2))^{1/2} + \left(\frac{\log(C_2/\varepsilon)}{C_3}\right)^{1/2}\right) \leq \frac{2\sqrt{2}M_\Phi}{bn^{1/2}}\left(C_1^2\log(2) + \frac{\log(C_2/\varepsilon)}{C_3}\right)^{1/2},$$

where the constants C_1, C_2, C_3 are the ones of Corollary 2.12. Easy manipulations give the result.

Proof of Lemma 2.26 Taking the supremum of each element in the sum we find that

$$|B_{n,1}(\varphi)| \leq \frac{M_\Phi}{b^8}\sup_{(u,x)\in\mathcal{K}}|H_h(u,x) - H(u,x)|\sup_{(u,x)\in\mathcal{K}}|\hat{H}^{(i)}_{0,n}(u,x)|.$$

An appeal to Lemma 2.22, Lemma 2.23 combined with eq. (2.52) gives the first result. Concering $B_{n,2}$, we write

$$|B_{n,2}(\varphi)| \leq \frac{M_\Phi}{b}\sup_{(t,x)\in\mathcal{K}}\left|\int_0^t \frac{c(u \mid x)}{H(u,x)}\left(H_{0,h}(\mathrm{d}u,x) - H_0(\mathrm{d}u,x)\right)\right|.$$

Because for any signed measure ν on $\mathbb{R}_+$ and any measurable function f with total variation at most 1 vanishing at infinity, we have (Dudley [1992]),

$$\left|\int f(u)\nu(\mathrm{d}u)\right| \leq \sup_{t\in\mathbb{R}}\left|\int_0^t \nu(\mathrm{d}u)\right|, \tag{2.61}$$

we conclude that

$$|B_{n,2}(\varphi)| \leq M\sup_{(u,x)\in\mathcal{K}}|H_{0,h}(u,x) - H_0(u,x)|.$$

where $M > 0$ depends only on L, b and M_Φ. Conclude by using the bound given in Lemma 2.22.

Proof of Lemma 2.27 Observe that, for $i \neq j$, we have

$$v_{i,j,j}(\varphi) = -\frac{\delta_i\varphi(T_i,X_i)}{S_C(T_i \mid X_i)}\mathbb{1}_{T_j\leq T_i}\frac{(1-\delta_j)K_{ij}c(T_j \mid X_i)}{H(T_j,X_i)^2}H_h(T_j,X_i).$$

It follows from eq. (2.60) that

$$\left|v_{i,j,j}(\varphi)\right| \leq \frac{M_\Phi}{b^8}\|K\|_\infty h^{-d}L,$$

and since $V''_{n,1}(\varphi)$ is a sum over $n(n-1)$ such terms divided by $n(n-1)^2$ we get the stated bound.

Proof of Lemma 2.28 We will use the expression

$$v_{i,j,k}(\varphi) = w_{i,j,k}(\varphi)K_{ik}K_{ij} - \mathbb{E}\left[w_{i,j,k}(\varphi)K_{ik}K_{ij} \mid Z_i, Z_j\right],$$

with

$$w_{i,j,k}(\varphi) = \frac{\delta_i \varphi(T_i, X_i)}{S_C(T_i \mid X_i)} \mathbb{1}_{T_j \leq \tilde{Y}_i} \frac{(1-\delta_j)c(T_j \mid X_i)}{H(T_j, X_i)^2} \mathbb{1}_{T_k > T_j}.$$

Using eq. (2.60), we have that

$$\left|w_{i,j,k}\right| \leq \frac{M_\Phi}{b^8}. \tag{2.62}$$

Recall that with probability 1,

$$\mathbb{E}\left[v_{i,j,k}(\varphi) \mid Z_i, Z_j\right] = \mathbb{E}\left[v_{i,j,k}(\varphi) \mid Z_i\right] = \mathbb{E}\left[v_{i,j,k}(\varphi) \mid Z_j\right] = 0.$$

As a result, the quantities $U_{n,1}^{(k)}(\varphi)$, $k \in \{1,2,3\}$ are degenerate U-statistics of degree 3, 2 and 2 respectively. For this reason we can apply Corollary 2.12 to each of them as soon as their respective kernels are shown to form VC classes. The kernel of $h^{2d}n(n-1)^2U_{n,1}^{(1)}$ is

$$h^{2d}\Big(v_{i,j,k}(\varphi) - \mathbb{E}\left[v_{i,j,k}(\varphi) \mid Z_j, Z_k\right] - \mathbb{E}\left[v_{i,j,k}(\varphi) \mid Z_i, Z_k\right] + \mathbb{E}\left[v_{i,j,k}(\varphi) \mid Z_k\right]\Big).$$

Lemma 2.14 and Corollary 17 in Nolan and D. Pollard (1987) implies that it is of VC type with constant envelope $8U$ as soon as $\{v_{i,j,k}(\varphi)\}$ is of VC type with envelope C defined as

$$U = \frac{M_\Phi}{b^8}\|K\|_\infty^2.$$

The later is true in virtue of Lemmas 2.13 and 2.16. The same arguments implies that the kernels of $h^{2d}n(n-1)U_{n,1}^{(2)}(\varphi)$ and $h^{2d}n(n-1)U_{n,1}^{(3)}(\varphi)$ are of VC type with the constant envelope $4C$. In what follows we specify, for each $U_{n,1}^{(k)}(\varphi)$, the value of σ to use in the application of Corollary 2.12.

The bound for $U_{n,1}^{(1)}(\varphi)$. Observe that

$$\mathbb{E}\left[\left(h^{2d}w_{i,j,k}(\varphi)K_{ik}K_{ij}\right)^2\right] \leq \left(\frac{M_\Phi}{b^8}\right)^2 \mathbb{E}\left[K\left(\frac{X_1 - X_2}{h}\right)^2 K\left(\frac{X_1 - X_3}{h}\right)^2\right] \leq \left(\frac{M_\Phi}{b^8}\right)^2 L^2 c_K^4 h^{2d},$$

where $c_K^2 = \int K^2(x)\,\mathrm{d}x$. Since we have a sum of 8 terms in the U-statistics of interest, $h^{2d}n(n-1)^2U_{n,1}^{(1)}(\varphi)$, each having an L_2-norm smaller that $\mathbb{E}[h^{4d}v_{i,j,k}(\varphi)^2]$ (by Jensen's inequality), we obtain a bound for the resulting variance (using Minkowski's inequality), of

$$\mathbb{V}\left[h^{2d}n(n-1)^2U_{n,1}^{(1)}(\varphi)\right] \leq 8^2\left(\frac{M_\Phi}{b^8}\right)^2 L^2c_K^4h^{2d}.$$

We apply Corollary 2.12 with $k=3$ and a value for σ^2 larger or equal than the previous bound. Note that $h \leq h_0$ and take

$$\sigma^2 = 8^2\left(\frac{M_\Phi}{b^8}\right)^2 L^2c_K^4h^{2d},$$
$$\|G\|_\infty = 8M_\Phi\frac{\|K\|_\infty^2}{b^8}.$$

Then, the conditions are

$$\frac{\|K\|_\infty^4}{L^2c_K^4h_0^d}\left(C_1^{2/3}\log\left(\frac{2\|K\|_\infty^2}{h^{d/2}h_0^{d/2}Lc_K^2}\right)+\frac{\log(C_2/\varepsilon)}{C_3}\right) \leq nh^d,$$
$$L^2c_K^4h^dh_0^d \leq \|K\|_\infty^4,$$

where C_1, C_2 and C_3 are the constants in Corollary 2.12. The latter conditions are indeed of the type $h \leq h_0$ and $nh^d \geq M_2|\log(\varepsilon h^{d/2})|$ which in as a result, gives

$$\sup_{\varphi\in\Phi}\left|h^{2d}n(n-1)^2U_{n,1}^{(1)}(\varphi)\right|$$
$$\leq \widetilde{M}_1h^{d/2}n^{3/2}\left(C_1\left(\log\left(\frac{\widetilde{M}_2}{h^{d/2}}\right)\right)^{3/2}+\left(\frac{\log(C_2/\varepsilon)}{C_3}\right)^{3/2}\right),$$

where $\widetilde{M}_1$ and $\widetilde{M}_2$ are constants depending on M_Φ, L, K, b, and h_0. To recover the stated result, one just needs to divide the previous bound by $n(n-1)(n-2)h^{2d}$ and to use similar manipulations as the ones presented at the end of the proof of Lemma 2.23.

The bound for $U_{n,1}^{(2)}(\varphi)$. In what follows, we use the shortcut

$$\mathbb{E}[\cdot \mid Z_i, Z_j] = \mathbb{E}[\cdot \mid i, j].$$

The kernel of $h^{2d}n(n-1)U_{n,1}^{(2)}(\varphi)$ is given by

$$h^{2d}\left(\mathbb{E}\left[v_{i,j,k}(\varphi) \mid j,k\right]-\mathbb{E}\left[v_{i,j,k}(\varphi) \mid k\right]\right)$$
$$= h^{2d}\Big(\mathbb{E}\left[w_{i,j,k}(\varphi)K_{ik}K_{ij} \mid j,k\right]-\mathbb{E}\left[w_{i,j,k}(\varphi)K_{ik}K_{ij} \mid j\right]$$
$$-\mathbb{E}\left[w_{i,j,k}(\varphi)K_{ik}K_{ij} \mid k\right]+\mathbb{E}\left[w_{i,j,k}(\varphi)K_{ik}K_{ij}\right]\Big).$$

By Jensen's inequality and Minkowski's inequality, the variance is then smaller than

$$4^2h^{4d}\mathbb{E}\left[\mathbb{E}\left[w_{i,j,k}(\varphi)K_{ik}K_{ij} \mid j,k\right]^2\right] \leq 4^2h^{4d}\left(\frac{M_\Phi}{b^8}\right)^2 \mathbb{E}\left[K_{ij}K_{ik} \mid j,k\right]^2.$$

But we have

$$\begin{aligned}\mathbb{E}\left[K_{ij}K_{ik} \mid j,k\right] &= \int K_h(x-X_j)K_h(x-X_k)g(x)\,\mathrm{d}x \\ &\leq L\int K(u)K_h(X_j-X_k+hu)\,\mathrm{d}u \\ &\leq LK_h^*(X_k-X_j),\end{aligned}$$

where

$$\begin{aligned}K^* &= K*K, \\ K_h^*(u) &= \frac{K^*(u/h)}{h^d}.\end{aligned}$$

Note that $\int K^*(u)\,\mathrm{d}u = 1$ and $\|K^*\|_\infty \leq \|K\|_\infty$. This previous equalities implie that

$$\begin{aligned}4^2h^{4d}\mathbb{E}\left[\mathbb{E}\left[w_{i,j,k}(\varphi)K_{ik}K_{ij} \mid j,k\right]^2\right] &\leq 4^2h^{4d}\left(\frac{M_\Phi L}{b^8}\right)^2 \mathbb{E}\left[K_{jk}^{*2}\right] \\ &\leq 4^2h^{3d}\left(\frac{M_\Phi L}{b^8}\right)^2 Lc_{K^*}^2,\end{aligned}$$

where $c_K^2 = \int K^2(x)\,\mathrm{d}x$. The bound eq. (2.56) is thus obtained by applying Corollary 2.12 to $h^{2d}n(n-1)U_{n,1}^{(2)}(\varphi)$ with $k=2$ and

$$\begin{aligned}\sigma^2 &= 4^2h^{2d}h_0^d\left(\frac{M_\Phi L}{b^8}\right)^2 Lc_{K^*}^2 \\ \|G\|_\infty &= 4\frac{M_\Phi}{b^8}\|K\|_\infty.\end{aligned}$$

The bound for $U_{n,1}^{(3)}(\varphi)$. Similar to the treatment of $U_{n,1}^{(2)}(\varphi)$, we apply Corollary 2.12 to $h^{2d}n(n-1)U_{n,1}^{(3)}(\varphi)$ with $k=2$. The kernel of $h^{2d}n(n-1)U_{n,1}^{(3)}(\varphi)$ is given by

$$\begin{aligned}&h^{2d}\left(\mathbb{E}\left[v_{i,j,k}(\varphi) \mid i,k\right] - \mathbb{E}\left[v_{i,j,k}(\varphi) \mid k\right]\right) \\ &\quad = h^{2d}\Big(\mathbb{E}\left[w_{i,j,k}(\varphi)K_{ik}K_{ij} \mid i,k\right] - \mathbb{E}\left[w_{i,j,k}(\varphi)K_{ik}K_{ij} \mid i\right] \\ &\qquad\qquad - \mathbb{E}\left[w_{i,j,k}(\varphi)K_{ik}K_{ij} \mid k\right] + \mathbb{E}\left[w_{i,j,k}(\varphi)K_{ik}K_{ij}\right]\Big).\end{aligned}$$

We then have the uniform bound

$$\|G\|_\infty = 4\frac{M_\Phi}{b^8}L\|K\|_\infty,$$

and variance bound

$$\begin{aligned}4^2h^{4d}\mathbb{E}\left[\mathbb{E}\left[w_{i,j,k}(\varphi)\mid i,k\right]^2\right] &\le 4^2h^{4d}\left(\frac{M_\Phi L}{b^8}\right)^2\mathbb{E}\left[K_{ik}^2\right]\\ &\le 4^2h^{3d}\left(\frac{M_\Phi L}{b^8}\right)^2 Lc_K^2.\end{aligned}$$

Proof of Lemma 2.29 Based on conditioning arguments, we have

$$\mathbb{E}\left[v_{i,j,k}(\varphi)\mid Z_k\right] = A_k(\varphi) - \mathbb{E}\left[A_k(\varphi)\right],$$

where $A_k(\varphi) = \mathbb{E}[w_{i,j,k}(\varphi)K_{ik}K_{ij}\mid k]$. We now show that the class of functions

$$\left\{Z_k \mapsto h^d A_k(\varphi) : \varphi \in \Phi\right\},$$

is a VC class with constant envelope. We first define

$$\begin{aligned}\beta_1(Z_i,Z_k) &= K_{ik}\frac{\delta_i\varphi(T_i,X_i)}{S_C(T_i\mid X_i)},\\ \beta_2(Z_i,Z_k) &= \int M(X_i+hu,Z_i,Z_k)K(u)\,\mathrm{d}u,\\ M(X_j,Z_i,Z_k) &= \int \frac{\mathbb{1}_{u\le\tilde{Y}_i,u<T_k}c(u\mid X_i)}{H(u,X_i)^2}H_0(\mathrm{d}u\mid X_j),\end{aligned}$$

and observe that

$$\begin{aligned}&\mathbb{E}\left[w_{i,j,k}(\varphi)K_{ik}K_{ij}\mid i,k\right]\\ &\quad=\beta_1(Z_i,Z_k)\mathbb{E}\left[\frac{\mathbb{1}_{T_j\le\tilde{Y}_i}(1-\delta_j)c(T_j\mid X_i)}{H(T_j,X_i)^2}\mathbb{1}_{T_k>T_j}K_{ij}\,\middle|\, i,k\right]\\ &\quad=\beta_1(Z_i,Z_k)\mathbb{E}\left[M(X_j,Z_i,Z_k)K_{ij}\mid i,k\right]\\ &\quad=\beta_1(Z_i,Z_k)\int M(X_i+hu,Z_i,Z_k)g(X_i+hu)K(u)\,\mathrm{d}u,\end{aligned}$$

where we have used Assumption 2.1 and the fact that $H_0(\mathrm{d}u|x) = S_Y(u-|x)S_C(\mathrm{d}u|x)$. Because for any f with total variation at most 1 vanishing at infinity, we have (Dudley [1992]),

$$\begin{aligned}\left|\int f(u)\left(H_0(\mathrm{d}u\mid x)-H_0(\mathrm{d}u\mid x')\right)\right| &\le \sup_{u\in\mathbb{R}}|H_0(u\mid x)-H_0(u\mid x')|\\ &\le L\|x-x'\|,\end{aligned}$$

where the last inequality is a consequence of Assumption 2.2. The same holds true for g in virtue of Assumption 2.2. Hence, the map M is uniformly Lipschitz with respect to X_j. Appealing to Lemma 2.15, we obtain that the kernel $h^d\mathbb{E}\left[w_{i,j,k}(\varphi)K_{ik}K_{ij} \mid i,k\right]$ is VC with constant envelope $\|K\|_\infty M_\Phi L/b^8$. The same holds true for $h^d A_k(\varphi)$ by Lemma 2.14. Moreover, observe that for all $(\varphi,h) \in \Phi\times]0,h_0]$, using eq. (2.62), we have almost surely,

$$|A_k(\varphi)| \le \frac{M_\Phi}{b^8}\mathbb{E}\left[K_{ij}K_{ik} \mid Z_k\right].$$

Because

$$\begin{aligned}
\mathbb{E}\left[K_{ij}K_{ik} \mid Z_k\right] &= \iint K_h(x-y)K_h(x-X_k)g(x)g(y)\,\mathrm{d}x\,\mathrm{d}y \\
&= \iint K(z)K_h(x-X_k)g(x)g(x-hz)\,\mathrm{d}x\,\mathrm{d}z \\
&\le L\int K_h(x-X_k)g(x)\underbrace{\left(\int K(z)\,\mathrm{d}z\right)}_{1}\mathrm{d}x \\
&\le L^2,
\end{aligned}$$

it follows that

$$\mathbb{E}\left[\left(h^d A_k(\varphi)\right)^2\right] \le h^{2d}\left(\frac{M_\Phi L^2}{b^8}\right)^2.$$

Applying Corollary 2.12 to the kernel $h^d(A_k(\varphi) - \mathbb{E}[A_k(\varphi)])$ with $k = 1$ and

$$\begin{aligned}
\|G\|_\infty &= 2\|K\|_\infty\frac{M_\Phi L}{b^8}, \\
\sigma^2 &= h^{2d}\left(\frac{M_\Phi L^2}{b^8}\right)^2,
\end{aligned}$$

yields the bound

$$\sup_{\varphi\in\Phi}\left|nh^d L_n'(\varphi)\right| \le \frac{M_\Phi L^2\sqrt{n}h^d}{b^8}\left(C_1\sqrt{\log(2)} + \sqrt{\log(C_2/\varepsilon)/C_3}\right),$$

with probability $1-\varepsilon$, provided a condition of the type

$$\begin{aligned}
nh^{2d} &\ge M_2\left|\log(h^d\varepsilon)\right|, \\
h &\le h_0.
\end{aligned}$$

Straightforward calculations then give the desired result.

Proof of Lemma 2.31 For all $\varphi \in \Phi$, we first set

$$w_{ij}(\varphi) = \frac{\delta_i \varphi(T_i, X_i) \mathbb{1}_{T_j \le T_i} (1-\delta_j) K_{ij} c(T_j \mid X_i)}{S_C(T_i \mid X_i) H(T_j \mid X_i)},$$

and observe next that

$$\begin{aligned}
V_{n,2}(\varphi) &= -\frac{1}{n} \sum_{i=1}^{n} \frac{\delta_i \varphi(T_i, X_i)}{S_C(T_i \mid X_i)} \int_0^{T_i} \frac{c(u \mid X_i)}{H(u, X_i)} \, \mathrm{d}\left(\hat{H}^{(i)}_{0,n}(u \, \mathrm{d}X_i) - H_{0,h}(u \, \mathrm{d}X_i)\right) \\
&= -\frac{1}{n(n-1)} \sum_{i \neq j} \frac{\delta_i \varphi(T_i, X_i)}{S_C(T_i \mid X_i)} \\
&\quad \times \left(\frac{\mathbb{1}_{T_j \le T_i} (1-\delta_j) K_{ij} c(T_j \mid X_i)}{H(T_j \mid X_i)} \right. \\
&\quad \left. - \int_0^{T_i} \frac{c(u \mid X_i)}{H(u, X_i)} \, \mathrm{d}H_{0,h}(u \, \mathrm{d}X_i) \right) \\
&= -\frac{1}{n(n-1)} \sum_{i \neq j} \Big(w_{ij}(\varphi) - \mathbb{E}\left[w_{ij}(\varphi) \mid Z_j \right] \Big) \\
&= U^{(1)}_{n,2}(\varphi) + U^{(2)}_{n,2}(\varphi),
\end{aligned}$$

where

$$\begin{aligned}
U^{(1)}_{n,2}(\varphi) = -\frac{1}{n(n-1)} \sum_{i \neq j} \Big(& w_{ij}(\varphi) - \mathbb{E}\left[w_{ij}(\varphi) \mid Z_j \right] \\
& - \mathbb{E}\left[w_{ij}(\varphi) \mid Z_i \right] + \mathbb{E}\left[w_{12}(\varphi) \right] \Big),
\end{aligned} \tag{2.63}$$

$$U^{(2)}_{n,2}(\varphi) = \frac{1}{n} \sum_{i=1}^{n} \Big(\mathbb{E}[w_{ij}(\varphi) \mid Z_i] - \mathbb{E}[w_{12}(\varphi)] \Big). \tag{2.64}$$

Hence, $V_{n,2}(\varphi)$ can be decomposed as the sum of a degenerate U-statistic defined in eq. (2.63) and an i.i.d. average defined in eq. (2.64). Note also that, by eq. (2.60), we have

$$\left| w_{ij}(\varphi) \right| \le \frac{M_\Phi}{b^5} K_{ij}.$$

The bound for $U^{(1)}_{n,2}(\varphi)$. By virtue of Lemmas 2.13 and 2.16, the collection of kernels of the degenerate U-statistics

$$\left\{ h^d n(n-1) U^{(1)}_{n,2}(\varphi) : (\varphi, h) \in \Phi \times]0, h_0] \right\},$$

forms a class of VC type with constants depending only on (υ, A), K and h_0. In addition, these terms are all bounded by $4M_\Phi \|K\|_\infty / b^5$ and we have:

$$\mathbb{V}\left[h^d w_{ij}(\varphi) \right] \le \left(\frac{4M_\Phi}{b^5} \right)^2 h^d L c_K^2.$$

It thus results from the application of Corollary 2.12 with $k = 2$ and

$$\sigma^2 = 4^2 \left(\frac{M_\Phi}{b^5}\right)^2 h^d L c_K^2,$$

that, with probability greater than $1 - \varepsilon$

$$\sup_{\varphi\in\Phi} \left| h^d n(n-1) U_{n,1}^{(2)}(\varphi) \right| \leq nh^{d/2} \widetilde{M}_1 \left(C_1 \log\left(\frac{\widetilde{M}_2}{h^{d/2}}\right) + \frac{\log(C_2/\varepsilon)}{C_3} \right), \quad (2.65)$$

where $\widetilde{M}_1$ and $\widetilde{M}_2$ depends on M_Φ, K, L, b, provided a condition of the type $h \leq h_0$ and $nh^d \geq M_2 |\log(\varepsilon h^{d/2})|$.

The bound for $U_{n,2}^{(2)}(\varphi)$. Following the proof of Lemma 2.29, the collection of kernels of $U_{n,2}^{(2)}(\varphi)$ is of VC type with constant envelope. Besides, we have, with probability one:

$$\mathbb{E}\left[w_{ij}(\varphi) \mid Z_i\right] \leq \frac{M_\Phi L}{b^5},$$

and therefore

$$\mathbb{V}\left[\mathbb{E}\left[w_{ij}(\varphi) \mid Z_i\right]\right] \leq \left(\frac{M_\Phi L}{b^5}\right)^2.$$

Thus, after applying Corollary 2.12 with $k = 1$ and

$$\sigma^2 = 4\left(\frac{M_\Phi L}{b^5}\right)^2,$$
$$\|G\|_\infty = 2\frac{M_\Phi L}{b^5},$$

we obtain that, with probability $1 - \varepsilon$,

$$\sup_{\varphi\in\Phi} \left| nU_{n,2}^{(2)}(\varphi) \right| \leq C\sqrt{n} \left(C_1 \sqrt{\log(2)} + \sqrt{\frac{\log(C_2/\varepsilon)}{C_3}} \right), \quad (2.66)$$

where C depends on M_Φ, K, L, b, provided a condition of the type $n \geq M_3 |\log(M_4/\varepsilon)|$ holds true but this is already implied by $h \leq h_0$ and $nh^d \geq M_2 |\log(\varepsilon h^{d/2})|$ whenever h_0 is small. The bound stated in the lemma results from rearranging the bounds of eqs. (2.65) and (2.66).

Proof of Lemma 2.32 Using the triangle inequality together with eqs. (2.52) and (2.53) and Lemmas 2.22 and 2.23, we get with probability $1-\varepsilon$ that

$$\max_{i=1,\dots n}\sup_{(t,x)\in\mathbb{K}}\left|\hat{H}^{(i)}_{0,n}(t,x)-H_0(t,x)\right|\leq \widetilde{M}_1\left(\frac{1}{nh^d}+\sqrt{\frac{|\log(\varepsilon h^{d/2})|}{nh^d}}+h^2\right),$$

$$\max_{i=1,\dots n}\sup_{(t,x)\in\mathbb{K}}\left|\hat{H}^{(i)}_{n}(t,x)-H(t,x)\right|\leq \widetilde{M}_2\left(\frac{1}{nh^d}+\sqrt{\frac{|\log(\varepsilon h^{d/2})|}{nh^d}}+h^2\right).$$

We suppose further that both previous inequalities are realized. Note that under the mentioned condition on (n,h), it holds that $\forall(t,x)\in\mathbb{K}$

$$\inf_{i=1,\dots,n}\hat{H}^{(i)}_n(t,x)\geq\frac{b^3}{2}.$$

In a similar fashion as in the proof of Theorem 2.6 (see eqs. (2.49) to (2.51)), we apply Lemma 2.19 to get that

$$\sup_{(t,x)\in\mathbb{K}}\left|\hat{S}^{(i)}_{C,n}(t\mid x)-S_C(t\mid x)\right|\leq\frac{2}{b}\sup_{(t,x)\in\mathbb{K}}\left|\hat{\Lambda}^{(i)}_{C,n}(t\mid x)-\Lambda_C(t\mid x)\right|.$$

Then, we apply Lemma 2.20, with $\theta_1=b^3$, $\theta_2=b^3/2$, $\beta=1$, to finally obtain that: $\forall i\in\{1,\dots,n\}$,

$$\begin{aligned}\sup_{(t,x)\in\mathbb{K}}\left|\hat{S}^{(i)}_{C,n}(t\mid x)-S_C(t\mid x)\right|&\leq\frac{2}{b}\left(\frac{2}{b^3}\sup_{(t,x)\in\mathbb{K}}\left|\hat{H}^{(i)}_{0,n}(t,x)-H_0(t,x)\right|\right.\\&\qquad\left.+\frac{2}{b^6}\sup_{(t,x)\in\mathbb{K}}\left|\hat{H}^{(i)}_{n}(t,x)-H(t,x)\right|\right)\\&\leq M_1\left(\frac{1}{nh^d}+\sqrt{\frac{|\log(\varepsilon h^{d/2})|}{nh^d}}+h^2\right).\end{aligned}$$

Hence, provided that $h\leq h_0$ and $nh^d\geq M_2|\log(\varepsilon h^{d/2})|$, we have

$$|R_n''(\varphi)|\leq\frac{2M_\Phi}{b^3}\sup_{(t,x)\in\mathbb{K}}\left|S_C(t\mid x)-\hat{S}^{(i)}_{C,n}(t\mid x)\right|^2.$$

Proof of Lemma 2.33 Recall first that

$$R_n'(\varphi)=-\frac{1}{n}\sum_{i=1}^n\frac{\delta_i\varphi(T_i,X_i)}{S_C(T_i\mid X_i)}\hat{b}^{(i)}_n(T_i\mid X_i),$$

where

$$\hat{b}_n^{(i)}(t \mid x) = -\int_0^t \frac{c(u \mid x)}{H(u,x)^2 \hat{H}_n^{(i)}(u,x)} \times \left(\hat{H}_n^{(i)}(u,x) - H(u,x)\right)^2 \hat{H}_{0,n}^{(i)}(\mathrm{d}u,x) - \int_0^t \frac{\hat{S}_{C,n}^{(i)}(u- \mid x) - S_C(u- \mid x)}{S_C(u \mid x)} \hat{\Delta}_n^{(i)}(\mathrm{d}u \mid x),$$

and

$$\hat{\Delta}_n^{(i)}(\mathrm{d}u \mid x) = \hat{\Lambda}_{C,n}^{(i)}(\mathrm{d}u|x) - \Lambda_C(\mathrm{d}u|x).$$

The following argument is based on Lemma 2.21, stated in §2.8.2. Note that, on the event $\mathcal{E}_n$, we have:

$$\begin{aligned}\left|\hat{b}_n^{(i)}(t \mid x)\right| &\leq \frac{2}{b^{10}} \int \left(\hat{H}_n^{(i)}(u,x) - H_h(u,x)\right)^2 \hat{H}_n^{(i)}(\mathrm{d}u,x) \\ &\quad + \left| \int_0^t \frac{\hat{S}_{C,n}^{(i)}(u- \mid x) - S_C(u- \mid x)}{S_C(u \mid x)} \hat{\Delta}_n^{(i)}(\mathrm{d}u \mid x) \right| \\ &\leq \frac{2}{b^{10}} \sup_{(u,x)\in\gamma_b} \left|\hat{H}_n^{(i)}(u,x) - H_h(u,x)\right|^2 \\ &\quad + \left| \int_0^t \frac{\hat{S}_{C,n}^{(i)}(u- \mid x) - S_C(u- \mid x)}{S_C(u \mid x)} \hat{\Delta}_n^{(i)}(\mathrm{d}u|x) \right|.\end{aligned}$$

The application of the Lemma 2.21, with $S^{(2)}(u) = S_C(u \mid x)$, $S^{(1)}(u) = \hat{S}_{C,n}^{(i)}(u \mid x)$, $\beta = 1$, $\theta = b$ and

$$\begin{aligned}\Lambda^{(1)}(u) &= \hat{\Lambda}_{C,n}^{(i)}(u \mid x) \\ &= -\int_{s=0}^u \frac{\hat{H}_{0,n}^{(i)}(\mathrm{d}s,x)}{\hat{H}_n^{(i)}(s-,x)}, \\ \Lambda^{(2)}(u) &= \Lambda_C(u \mid x) \\ &= -\int_{s=0}^u \frac{H_0(\mathrm{d}s,x)}{H(s-,x)},\end{aligned}$$

yields,

$$\begin{aligned}&\frac{1}{C} \left| \int_0^t \frac{\left(\hat{S}_{C,n}^{(i)}(u- \mid x) - S_C(u- \mid x)\right)}{S_C(u \mid x)} \hat{\Delta}_n^{(i)}(\mathrm{d}u \mid x) \right| \\ &\leq \sup_{(u,x)\in\gamma_b} \left|\hat{H}_n^{(i)}(u,x) - H(u,x)\right|^2 + \sup_{(u,x)\in\gamma_b} \left|\hat{H}_{0,n}^{(i)}(u,x) - H_0(u,x)\right|^2 \\ &\qquad + \sup_{(u,x)\in\gamma_b} \left|\hat{W}_n^{(i)}(u,x)\right|,\end{aligned}$$

where $C > 0$ depends on b and $\hat{W}_n^{(i)}(t,x)$ is defined as

$$\int_0^t \int_0^u c(s \mid x) \frac{\left(\hat{H}_{0,n}^{(i)}(\mathrm{d}s,x) - H_0(\mathrm{d}s,x)\right)}{H(s,x)} \frac{\left(\hat{H}_{0,n}^{(i)}(\mathrm{d}u,x) - H_0(\mathrm{d}u,x)\right)}{S_C(u \mid x)H(u,x)}.$$

Using eqs. (2.52) and (2.53) combined with Lemmas 2.22 and 2.23, we obtain that with probability at least $1-\varepsilon$:

$$\sup_{(u,x)\in\gamma_b} \left|\hat{H}_n^{(i)}(u,x) - H(u,x)\right|^2 + \sup_{(u,x)\in\gamma_b} \left|\hat{H}_{0,n}^{(i)}(u,x) - H_0(u,x)\right|^2 \leq M_1 \left(\frac{1}{(nh^d)^2} + \frac{\left|\log(\varepsilon h^{d/2})\right|}{nh^d} + h^4 \right),$$

as soon as $h \leq h_0$ and $M_2|\log(\varepsilon h^{d/2})| \leq nh^d$. It remains to show that, with probability at least $1-\varepsilon$:

$$\max_{i\in\{1,\ldots,n\}} \sup_{(u,x)\in\gamma_b} \left|\hat{W}_n^{(i)}(u,x)\right| \leq M_1 \left(\frac{\left|\log(h^{d/2}\varepsilon)\right|}{nh^d} + h^2 \right), \tag{2.67}$$

as soon as $h \leq h_0$ and $M_2|\log(\varepsilon h^{d/2})| \leq nh^d$. We first define $\hat{W}_{n,1}(t,x)$, for all $(t,x) \in \mathbb{K}$,

$$\int_0^t \int_0^u c(s \mid x) \frac{\hat{H}_{0,n}(\mathrm{d}s,x) - H_0(\mathrm{d}s,x)}{H(s,x)} \frac{\hat{H}_{0,n}(\mathrm{d}u,x) - H_0(\mathrm{d}u,x)}{S_C(u \mid x)H(u,x)},$$

and notice that, since

$$\frac{c(s \mid x)}{H(s,x)S_C(u \mid x)H(u,x)} \leq \frac{1}{b^8},$$

by using eq. (2.60), we have by virtue of eq. (2.52)

$$\max_{i\in\{1,\ldots,n\}} \sup_{(t,x)\in\mathbb{K}} \left|\hat{W}_{n,1}(t,x) - \hat{W}_n^{(i)}(t,x)\right| \leq \frac{C}{nh^d},$$

where C is a constant depending on b and K only. Let

$$\alpha_1(u,x) = \frac{1}{S_C(u \mid x)H(u,x)} \int_0^u c(s \mid x) \frac{H_{0,h}(\mathrm{d}s,x) - H_0(\mathrm{d}s,x)}{H(s,x)},$$

and note that

$$\sup_{(u,x)\in\mathcal{K}} |\alpha_1(u,x)| \leq M_1 \sup_{(u,x)\in\mathcal{K}} |H_{0,h}(u,x) - H_0(u,x)| \leq M_2 h^2.$$

We define $\hat{W}_{n,2}(t,x)$ as

$$\int_0^t \int_0^u c(s \mid x) \frac{\hat{H}_{0,n}(\mathrm{d}s,x) - H_{0,h}(\mathrm{d}s,x)}{H(s,x)} \frac{\hat{H}_{0,n}(\mathrm{d}u,x) - H_{0,h}(\mathrm{d}u,x)}{S_C(u \mid x)H(u,x)},$$

we have

$$\begin{aligned}
\hat{W}_n(t,x) &= \hat{W}_{n,2}(t,x) \\
&+ \int_0^t \int_0^u c(s \mid x) \frac{\hat{H}_{0,n}(\mathrm{d}s,x) - H_{0,h}(\mathrm{d}s,x)}{H(s,x)} \frac{H_{0,h}(\mathrm{d}u,x) - H_0(\mathrm{d}u,x)}{S_C(u \mid x)H(u,x)} \\
&\quad + \int_0^t \alpha_1(u,x) \left(\hat{H}_{0,n}(\mathrm{d}u,x) - H_0(\mathrm{d}u,x)\right) \\
&\quad + \int_0^t \alpha_1(u,x) \left(H_{0,h}(\mathrm{d}u,x) - H_0(\mathrm{d}u,x)\right).
\end{aligned}$$

Applying Fubini's theorem in the second term, we see that the last three terms are similar. We give the details only for the second one. We have

$$\begin{aligned}
&\left| \int_0^t \alpha_1(u,x) \left(\hat{H}_{0,n}(\mathrm{d}u,x) - H_0(\mathrm{d}u,x)\right) \right| \\
&\quad \le \int_0^t |\alpha_1(u,x)| \left(\hat{H}_{0,n}(\mathrm{d}u,x) + H_0(\mathrm{d}u,x)\right) \\
&\quad \le \sup_{(u,x)\in\mathcal{K}} |\alpha_1(u,x)| \sup_{(u,x)\in\mathcal{K}} \left|\hat{H}_{0,n}(u,x) + H_0(u,x)\right| \\
&\quad \le M_2 h^2 \sup_{(u,x)\in\mathcal{K}} \left|\hat{H}_{0,n}(u,x) + H_0(u,x)\right|.
\end{aligned}$$

In addition, observe that

$$\begin{aligned}
\hat{W}_{n,2}(t,x) &= \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n v_{ij}(t,x) \\
&= \frac{1}{n^2} \sum_{i\neq j} v_{ij}(t,x) + \frac{1}{n^2} \sum_{i=1}^n v_{ii}(t,x) \\
&\stackrel{\text{def}}{=} U_n(t,x) + M_n(t,x), \qquad (2.68)
\end{aligned}$$

where, for all $1 \le i, j \le n$, we set

$$\begin{aligned}
v_{ij}(t,x) &= u_{ij}(t,x) - \mathbb{E}\left[u_{ij}(t,x) \mid Z_i\right] - \mathbb{E}\left[u_{ij}(t,x) \mid Z_j\right] + \mathbb{E}\left[u_{1,2}(t,x)\right], \\
u_{ij}(t,x) &= \xi_{i,j}(x) \mathbb{1}_{T_i \le t} K_h(X_i - x) K_h(X_j - x), \\
\xi_{i,j}(x) &= \frac{\delta_i \delta_j c(T_j \mid x)}{S_C(T_i,x)H(T_i,x)H(T_j,x)} \mathbb{1}_{T_j \le T_i}.
\end{aligned}$$

Because we have, for all $(t,x) \in \mathbb{K}$,

$$\mathbb{E}\left[v_{12}(t,x) \mid Z_1\right] = \mathbb{E}\left[v_{12}(t,x) \mid Z_2\right] = 0,$$

the collection of random variables

$$\left\{n^2 h^{2d} U_n(t,x) : (t,x,h) \in \mathbb{K}\times]0,h_0]\right\},$$

is a degenerate U-process of order 2. The related class of kernels is uniformly bounded by $4||K||_\infty^2/b^8$ and of VC type, by virtue of classic permanence properties recalled in §2.8.1. Observe in addition that

$$\begin{aligned}\mathbb{V}\left[h^{2d} v_{12}(t,x)\right] &\le h^{4d} 4^2 \mathbb{E}\left[u_{12}^2(t,x)\right] \\ &\le h^{4d}\left(\frac{4}{b^8}\right)^2 \mathbb{E}\left[K_{1x}^2 K_{2x}^2\right] \\ &\le h^{2d}\left(\frac{4}{b^8}\right)^2 L^2 c_K^4.\end{aligned}$$

Applying Corollary 2.12 with $k = 2$ and

$$\begin{aligned}\sigma^2 &= 4h^d h_0^d \frac{L^2}{b^8} c_K^4, \\ \|G\|_\infty &= 4\frac{\|K\|_\infty^2}{9b^8},\end{aligned}$$

we obtain that, with probability greater than $1-\varepsilon$,

$$\sup_{(t,x)\in\mathbb{K}} |U_n(t,x)| \le M_1 \frac{\left|\log(\varepsilon h^{d/2})\right|}{nh^d}, \tag{2.69}$$

as soon as $M_2|\log(\varepsilon h^{d/2})| \le nh^d$ and $h \le h_0$. Notice now that, for all $(t,x) \in \mathbb{K}$,

$$M_n(t,x) = L_n(t,x) + R_n(t,x),$$

where L_n and R_n are defined by

$$\begin{aligned}L_n(t,x) &\stackrel{\text{def}}{=} \frac{1}{n^2}\sum_{i=1}^n (v_{ii}(t,x) - \mathbb{E}\left[v_{11}(t,x)\right]), \\ R_n(t,x) &\stackrel{\text{def}}{=} \frac{1}{n}\mathbb{E}\left[v_{11}(t,x)\right].\end{aligned}$$

Observing that, for all $(t,x) \in \mathbb{K}$, we have

$$\begin{aligned}
\left|h^{2d} v_{11}(t,x)\right| &\leq 4\frac{\|K\|_\infty^2}{b^8}, \\
\mathbb{V}\left[h^{2d} v_{11}(t,x)\right] &\leq h^{4d} 4^2 \mathbb{E}\left[u_{11}^2(t,x)\right] \\
&\leq h^{4d}\left(\frac{4}{b^8}\right)^2 \mathbb{E}\left[K_{1x}^4\right] \\
&\leq h^d \left(\frac{4}{b^8}\right)^2 L \int K^4(x)\,\mathrm{d}x.
\end{aligned}$$

We can apply Corollary 2.12 with $k = 1$ to the empirical sums

$$\left\{n^2 h^{2d} L_n(t,x) : (t,x,h) \in \mathbb{K}\times]0,h_0]\right\},$$

which gives us, with probability at least $1-\varepsilon$,

$$\sup_{(t,x)\in\mathbb{K}} |L_n(t,x)| \leq \widetilde{M}_1 \frac{\sqrt{\left|\log(\widetilde{M}_2/h^{d/2})\right|} + \sqrt{\log(C_2/\varepsilon)/C_3}}{(nh^d)^{3/2}}, \tag{2.70}$$

where $\widetilde{M}_1$ and $\widetilde{M}_2$ are constants depending on K, b and L. The previous bound is valid whenever $M_2|\log(\varepsilon h^{d/2})| \leq nh^d$ and $h \leq h_0$. We also have

$$\frac{1}{n}\mathbb{E}\left[|v_{11}(t,x)|\right] \leq \frac{4}{n}\mathbb{E}\left[|u_{11}(t,x)|\right] \leq \frac{4Lc_K^2}{b^8 nh^d}$$

This leads to the stated results.

Flexible Generative Survival Models | 3

3.1 Introduction

In the previous chapter, we showed how to adapt the ERM framework to the survival analysis setting; this approach, however, presupposes the ability to properly estimate the survival function of the censoring variable. While we proved empirically that a straightforward kernel estimator is enough to obtain good results, the quality of the resulting estimator depends heavily on the quality of the Kaplan-Meier weights i.e. of the estimator $1/\hat{S}_C$ of the true weights $1/\hat{S}_C$. Estimators of the IPCW weights can be constructed fairly easily from plugin estimators of S_C by reusing tools from the density estimation literature in order to learn a parametric family of the unobserved time C. This is the approach we adopted in the previous chapter by employing a kernel conditional formulation inspired by kernel density estimation.[71]

We will momentarily switch to the point of view of the estimation of S_Y as, due to the symmetry between Y and C, it is strictly identical to that of S_C but has a wider and more understandable applicability in general, barring our very specific need of S_C for the construction of IPCW weights. The fact that, in the censored setting, we only observe the variables (T, δ, X) instead of (Y, X) does not prove to a problem if the task is the estimation of the density. Notice, after simple manipulations, that it is possible in the right censored setting to write the likelihood of the observed data in terms of relevant quantities only:

$$\mathbb{P}(T, \delta \mid X) \propto p_Y(T \mid X)^{\delta} S_Y(T \mid X)^{1-\delta}, \tag{3.1}$$

where p_Y is the density of the unknown conditional distribution of $Y \mid X$, our unobserved variable of interest, and S_Y its corresponding survival function. From eq. (3.1) it is easy to see that the task of estimating the best, for the specific criterium of the log-likelihood, family defined by p_Y[72] can itself be framed as an empirical risk minimization problem[73]

$$\underset{p_Y, S_Y}{\operatorname{argmin}} \sum_{i=1}^{n} -\delta_i p_Y(T_i \mid X_i) - (1 - \delta_i) S_Y(T_i \mid X_i), \tag{3.2}$$

71: More generally known as a *Stone type* estimator.

72: And therefore $S_Y = \int p_Y \, \mathrm{d}t$.

73: Without any of the problems encountered in the previous chapter. There is no reweighing needed here.

and therefore solved as long as the estimators chosen for p_Y and S_Y result in a problem amenable to optimization.[74] Even without any parametric restriction on the choices of p_Y and S_Y, the previous quantity can be minimized unconditionally such that the nonparametric minimizer of the previous quantity is exactly the Kaplan-Meier estimator of Kaplan and Meier (1958).[75] It is then possible, given a sufficiently well behaved yet powerful parametric family of survival distributions defined by $(p_Y(\cdot \mid \theta), S_Y(\cdot \mid \theta))$, to learn a flexible estimator of (p_Y, S_Y). Note that while we refer to the tuple (p_Y, S_Y), which summarizes the quantities needed to compute eq. (3.2), we only need to estimate a single one of these quantities and then deduce the other by using the former as a plugin. Consequently, in this chapter we present a novel way to build a flexible estimator of the survival S by making use of normalizing flows. Through the use of a more expressive family of neural based estimators, we are able to obtain better conditional estimators of S_Y (resp. S_C) when the data generating process cannot be easily modelled using classical tools, as is usually the case for highly unstructured covariates such text, or when the usual model assumptions, such as for example the proportional hazards assumption, are violated.

Parts of this chapter generally follow the work published in Ausset, Ciffreo, et al. (2021) at IEEE DSAA'21, where we proposed the use of normalizing flows for survival analysis. In §3.3 we motivate the need for generative models for many applications including medicine and finance. In §3.3.2, in order to lay the groundwork for our method, we present normalizing flows as introduced by Rezende and Mohamed (2015) in their discrete formulation and by R. T. Q. Chen et al. (2018) in their continuous form. In §3.4 we show how normalizing flows can be adapted to the unconditional survival analysis setting while we reintroduce conditioning in §3.5 with conditional survival flows. Finally, in §3.6 we illustrate empirically how the flexibility of survival flows helps improve performance on common tasks while applications to finance will be studied in more details in chapter 5.

About this Chapter

The rest of this chapter is in large part reproduced from the paper "Individual Survival Curves with Conditional Normalizing Flows" with Tom Ciffreo, Timothée Papin, Stéphan Clémençon and François Portier presented at IEEE DSAA'21. Some of the results established in this chapter have applied internally at BNP Paribas on internal data. While impossible to reproduce those results in the thesis, similar experiments on related financial data while be presented later.

74: Ideally we would like the quantity in eq. (3.2) to be a sum of smooth convex functions, in practice we only ask for differentiability.

75: The proof actually makes use of the IPCW representation of the previous chapter: notice that both $\hat{S}_C$ and $\hat{S}_Y$ admit a coupled IPCW form in term of the other. By substituting in eq. (3.2), one obtains the Kaplan-Meier estimator.

Guillaume Ausset, Tom Ciffreo, et al. (2021). "Individual Survival Curves with Conditional Normalizing Flows". In: *DSAA'21*. IEEE International Conference on Data Science and Advanced Analytics

3.2 Estimators of the Survival

Multiple quantities can be equivalently modelled in order to uniquely define the distribution of the survival times, each leading to a different approach and therefore estimator of S_Y. The particular structure of the right censored problem enables the possibility to model interchangeably either the density $p_Y(\cdot \mid X = x)$ of $Y \mid X = x$, its survival function $S_Y(\cdot \mid X = x)$ defined by

$$S_Y(y \mid X = x) = \int_y^\infty p_Y(u \mid X = x)\,\mathrm{d}u,$$

the instantaneous hazard rate $\lambda_Y(\cdot \mid X = x)$ or the integrated hazard rate $\Lambda_Y(\cdot \mid X = x)$, defined by:

$$\begin{aligned}
\lambda_Y(y \mid X = x) &= \lim_{\mathrm{d}y \to 0} \frac{\mathbb{P}(y \le Y \le y + \mathrm{d}y \mid y \le Y, X = x)}{\mathrm{d}y}, \\
\Lambda_Y(y \mid X = x) &= \int_0^y \lambda_Y(u \mid X = x)\,\mathrm{d}u, \\
\Lambda_Y(y \mid X = x) &= \int_0^y \frac{F_Y(\mathrm{d}u)}{S_Y(u- \mid X)}.
\end{aligned} \tag{3.3}$$

Any of the previous four quantities fully characterizes the conditional law of Y given X and can be used to recover the other three through the identities

$$\begin{aligned}
\frac{\mathrm{d}S_Y(y \mid X = x)}{\mathrm{d}t} &= -\lambda_Y(y \mid X = x) S_Y(y \mid X = x), \\
S_Y(y \mid X = x) &= \exp\big(-\Lambda_Y(y \mid X = x)\big).
\end{aligned} \tag{3.4}$$

The links between all those quantities are summarized in fig. 3.1. The symmetry between the roles of Λ and S is more than apparent and follows from the product-integral construction of the survival. If one defines the interval measure in terms of hazard and survival, that is

$$\begin{aligned}
\Lambda_Y(s,t) &= \Lambda_Y\left(]s,t]\right) = \Lambda_Y(t) - \Lambda_Y(s), \\
S_Y(s,t) &= S_Y\left(]s,t]\right) = \frac{S_Y(t)}{S_Y(s)} = \mathbb{P}\left(Y > t \mid Y > s\right),
\end{aligned} \tag{3.5}$$

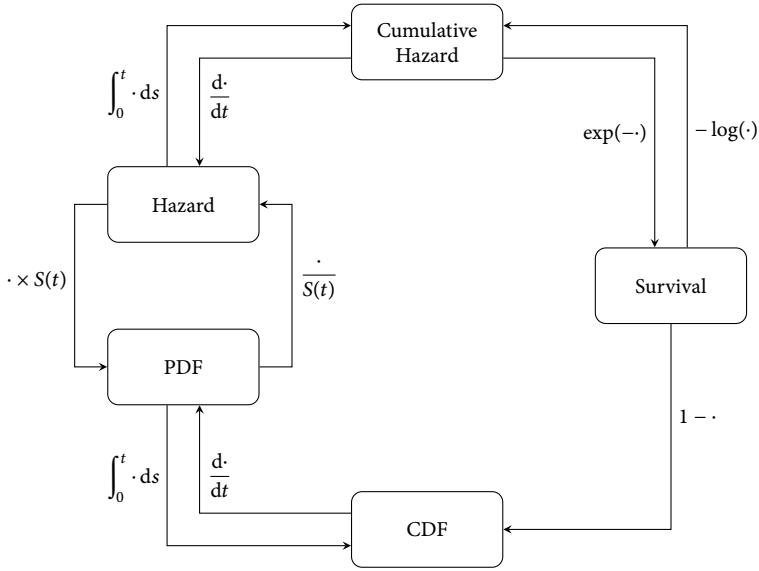


FIGURE 3.1: A map of the quantities defining the survival.

then Λ_Y is an additive interval function i.e. for $s \leq t \leq u$ we have

$$\Lambda_Y(s,u) = \Lambda_Y(s,t) + \Lambda_Y(t,u),$$

and S_Y is a multiplicative interval function i.e.

$$S_Y(s,u) = S_Y(s,t)S_Y(t,u),$$

The link between the two quantities can then be expressed infinitesimally as

$$\Lambda_Y(ds) = 1 - S_Y(ds),$$
$$S_Y(ds) = 1 - \Lambda_Y(ds),$$

and expressed in integral form as

$$\Lambda_Y(t) = \Lambda_Y(0,t) = \int_0^t (1 - S_Y(ds)),$$
$$S_Y(t) = S_Y(0,t) = \prod_0^t (1 - \Lambda_Y(ds)),$$

where $\prod$ is the product-integral[76] operator, defined analogously to the usual sum-integral as the limit of the product over partitions of $]0,t]$ with the mesh converging to zero.

The instantaneous hazard rate is usually considered the most natural quantity insofar as it can be directly interpreted: it represents the instantaneous probability of an event happening now as opposed to the density

76: The notion of product-integral was introduced in 1887 by Volterra (of the Lotka-Volterra's fame) to characterise forward-backward differential equations. As survival analysis, seen through the scope of counting processes, involves markov processes; the product-integral formulation plays an important role. The symbol here was introduced by Gill in his studies of the product-integral in the survival analysis setting (see Gill (1994); Andersen et al. (1993)), based on the earlier works of Nerney (1963) and Dobrushin (1953). Это советская власть до самого конца.

p_Y which represents the instantaneous probability of an event happening seen from the origin $Y = 0$. Moreover, the intensity as seen as an interval function as done in eq. (3.5), is additive and has a direct interpretation in terms of transition intensity when the survival process is seen as a markovian process. Choosing to estimate λ_Y is therefore a sensible choice. The simplest, unconditional and nonparametric, estimator of λ_Y is due to the work of Nelson (1969, 1972) and Aalen (1978) where the intensity is estimated by the ratio

$$\hat{\lambda}_Y(t_i) = \frac{d_i}{r_i},$$

where d_i is the number of events at time t_i and r_i is the number of individuals still alive *just before* t_i, also called the *risk set*. From this estimation one can then obtain the cumulative hazard from the hazard as a plugin:

$$\hat{\Lambda}_Y(t) = \sum_{i:t_i \leq t} \frac{d_i}{r_i}.$$

Similarly as done in the previous chapter, it is possible to kernelize the previous estimator in order to obtain a conditional version[77] at $X = x$

$$\hat{\Lambda}_Y(t \mid X = x) = \sum_{i:t_i \leq t} \frac{\sum K_h(X_i - x)\mathbb{1}_{T_i \leq u, \delta_i = 1}}{\sum K_h(X_j - x)\mathbb{1}_{T_j > u}}.$$

77: Surprisingly, to the best of my knowledge, this estimator do not appear to have any name.

A version of this estimator is for example studied by Dabrowska (1988) but, generally, those kernelized estimators are rarely used alone as they are highly susceptible to the pitfalls of the curse of dimensionality and therefore prone to overfitting and low performance.

Beyond nonparametric approaches, semi-parametric or fully parametric models of the hazards have enjoyed great successes. Under a proportionality hypothesis of the type $\lambda_Y(t \mid X) = \lambda_0(t)\lambda_{Y,x}(X)$, i.e. such that

$$\frac{\lambda_Y(t \mid X_i)}{\lambda_Y(t \mid X_j)} = \frac{\lambda_{Y,x}(X_i)}{\lambda_{Y,x}(X_j)},$$

where the last equality is constant with respect to the time. It is possible to form the partial log-likelihood

$$\sum_i \delta_i \left(\log(\lambda_{Y,x}(X_i)) - \log\left(\sum_{j:T_j \geq T_i} \lambda_{Y,x}(X_j) \right) \right), \tag{3.6}$$

which can be maximized straightforwardly. The usual choice of

$$\lambda_{Y,x}(X) = \exp(X^\intercal \beta),$$

yields the standard estimator of D. R. Cox (1972), which enjoys a lasting popularity due to its high interpretability: the coefficients β directly model the log odds ratios. As the choice of $\lambda_{Y,x}$ is left to the practitioners recent approaches, such as DeepSurv (Katzman et al. [2018]) where $\lambda_{Y,x}$ is taken as a deep neural network, have employed increasingly flexible families of functions. Of course, it is only possible at first glance from the previous quantity of eq. (3.6) to learn the proportional components of the hazard, which in many cases such as comparing the relative survival of one group compared to another is sufficient,[78] but in our case it is necessary to also estimate the baseline hazard λ_0. Fortunately, this is possible both in the parametric case where the distribution of the baseline survival is enforced[79] but also without specifying any model on λ_0. This latter, rather surprising, result is due to Breslow (1975) and fit in the more general non-parametric maximum likelihood estimation (NPMLE) framework. Breslow proposes the joint maximization of the likelihood in both $\lambda_{Y,x}$ and λ_0 by treating λ_0 as a piecewise constant between uncensored failure times. By maximizing

$$\sum_i \left(\delta_i \left(\log(\lambda_{Y,x}(X_i)) + \log(\lambda_0(T_i)) \right) - \int_0^{T_i} \lambda_{Y,x}(X_i)\lambda_0(u)\,\mathrm{d}u \right),$$

one obtains the surprisingly simple result that

$$\hat{\Lambda}_0(t) = \sum_{i=1}^{n} \frac{\mathbb{1}_{T_i \le t}\delta_i}{\sum_{j:T_j \ge T_i} \lambda_{Y,x}(X_j)}.$$

While the proportional hazard hypothesis is restrictive it is possible to relax it, either by allowing $\lambda_{Y,x}$ to depends on the time t (but still keeping the baseline hazard λ_0), which is the approach taken by Kvamme, Borgan, and Scheel (2019), or by modelling the survival as a mixture of K Cox models which is the approach taken by Nagpal et al. (2021) where they model the Z-integrated log-likelihood as

$$\sum_i \sum_{k=1}^{K} p_k(X_i) \Big(\delta_i \left(\log(\lambda_{Y,x,k}(X_i)) + \log(\lambda_{Y,0,k}(T_i)) \right) - \int_0^{T_i} \lambda_{Y,x,k}(X_i)\lambda_{Y,0,k}(u)\,\mathrm{d}u \Big), \quad (3.7)$$

where $p_k(X_i) = \mathbb{P}(Z = k \mid X = X_i)$ is the assignment probability of the k-th component with Z a latent variable introduced purely for the sake of making the problem easier. Note that the previous quantity of eq. (3.7) is not the log-likelihood $\mathcal{L}$ but the integrated log-likelihood $\mathbb{E}_{Z|X}[\mathcal{L}]$. In order to maximize the likelihood, it is necessary to use the expectation

78: Take for example the case of the study of a new medication compared to another, or a placebo. The *overall* efficacy of the medication may not be that important but only its performance compared to the existing medication.

79: As a way to regularize or to introduce priors, if there are reasons to believe some specific distribution is adapted.

80: It is possible to see survival analysis as a missing data problem, which can then be solved by use of EM algorithm. The use of latent variable and the expectation-maximization procedure is general enough to be of interest to all readers.

maximization (EM)[80] algorithm of Dempster, Laird, and D. B. Rubin (1977), that is to maximize a lower bound of the true log-likelihood by iteratively forming the expectation and then maximizing the newly obtained bound. While, as previously, it is possible to parametrize the mixtures themselves using neural networks as is done in DeepSurv; Nagpal et al. (2021) chose to parametrize the mixture parameters p_k themselves a a neural network of the form

$$p_k(x) = \text{softmax}(\text{nn}(x)).$$

Note that under the mixture model the hazard rates then have the form

$$\lambda_Y(\cdot \mid x) = \frac{\sum_k \mathbb{P}(t \mid x, Z = k)\mathbb{P}(Z = k \mid x)}{\sum_k S_Y(t \mid x, Z = k)\mathbb{P}(Z = k \mid x)},$$

which violates in general the proportional hazard assumption. As a consequence, it is not possible to directly maximize the partial likelihood as is done in the usual Cox model. Similarly in spirit to the previous approaches, Fernandez, Rivera, and Teh (2016) make use of Gaussian processes to model the instantaneous hazard rate such that

$$\begin{aligned} l(\cdot) &\sim \mathcal{GP}(0, K), \\ \lambda(t, X_i) &= \lambda_0(t)\sigma\left(l(t, X_i)\right), \end{aligned}$$

where $\mathcal{GP}$ denotes a Gaussian process and K is a kernel. Deviating from the proportional hazard hypothesis and related approaches, it is possible to consider an AFT formulation of the hazard,

$$\lambda_Y\left(t \mid X\right) = f\left(X\right)\lambda_{Y,0}\left(f(X)t\right).$$

It is possible to show that, under this model, both the density and the survival follow a simple relation

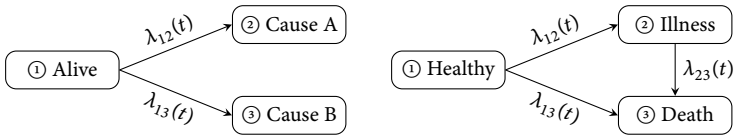
$$\begin{aligned} p_Y\left(t \mid X\right) &= f\left(X\right)p_{Y,0}\left(f(X)t\right), \\ S_Y\left(t \mid X\right) &= S_{Y,0}\left(f(X)t\right). \end{aligned}$$

Therefore, if we denote by ϵ the random variable distributed according to $p_{Y,0}$,[81] we can then formulate the time-to-event as a simple transformation of ϵ of the form

$$\begin{aligned} \log(Y) &= -\log\big(f(X)\big) + \log\big(Yf(X)\big) \\ &\stackrel{\text{def}}{=} -\log\big(f(X)\big) + \epsilon, \end{aligned} \tag{3.8}$$

More generally, and similarly to the approach we will introduce later, it is possible to entirely discard the proportional hazard or AFT hypothesis provided one is able to solve a differential equation, which is the approach

81: Note therefore, that this model can be seen simultaneously as a model of the hazard h_Y and therefore trained as usual by maximixation of the partial log-likelihood, but also as a regression model of the transformed target $\log(Y)$ and a transformed model of Y. The last point of view will be of particular interest to us later.



(A) Competing Hazards Model (B) Illness-Death Model

FIGURE 3.2: Multistate survival models.

taken by Groha, Schmon, and Gusev (2020). Remember that the partial likelihood can be written as

$$\sum_i \left(\delta_i \log(\lambda_Y(T_i, X_i)) - \int_0^{T_i} \lambda_Y(u, X_i)\, \mathrm{d}u \right).$$

By modelling $\lambda_Y(x, x)$ directly, for example as a neural network, it is then possible to maximize the previous quantity provided one is able to compute the integral of $\lambda_Y(x, x)$ as well as maximize objects involving integrally defined quantities. The previous approach has the advantage of imposing next to no modelling constraints such as the positivity or monotony constraints present in Λ_Y, S_Y or even p_Y.

Note that we earlier justified seeing λ as the natural quantity of the survival because of its interpretation as the intensity of a markovian process. It is actually possible to extend the survival models based on the estimation of λ_Y to multi-state survival analysis that is, instead of only considering the states alive/dead,[82] to also consider intermediate or adjacent states. This extended framework includes for example the competing hazards model of fig. 3.2a where death can occur for multiple reasons[83] or the illness-death model of fig. 3.2b. In multistate models, if we denote the state transition probabilities by

$$\mathbb{P}_{ij}(s, t) = \mathbb{P}_{ij}\left(Z(t) = j \mid Z(s) = i\right),$$

where Z is the state, with finite state space of size m and t_j^i the j-th time for the observation i, then the likelihood of the model can be written as

$$\prod_{i=1}^{n} \mathbb{P}\left(Z_i(t_1^i)\right) \prod_{j=2}^{m_i - 1} \mathbb{P}_{Z_i(t_{j-1}^i) Z_i(t_j^i)}(t_{j-1}^i, t_j^i) \lambda_{Z_i(t_{j-1}^i) Z_i(t_j^i)}(t_j^i)$$
$$\times \mathbb{P}_{Z_i(t_{m_i-1}^i) Z_i(t_{m_i}^i)}\left(t_{m_i-1}^i, t_{m_i}^i\right) \left(\lambda_{Z_i(t_{m_i-1}^i) Z_i(t_{m_i}^i)}(t_{m_i}^i)\right)^{\delta_i},$$

As we know that, under the Markov assumption, the transition probabilities must obey the Kolmogorov forward equation

$$\frac{\mathrm{d}\mathbb{P}_{ij}(s, t)}{\mathrm{d}t} = \sum_k \mathbb{P}_{ik}(s, t) \lambda_{kj}(t),$$

82: Referred as the 1-state model, if we consider "alive" to be a default state.

83: In this framework, each cause of death censor the other causes of deaths in a certain way. The usual censoring can also be present.

then, provided that we know how to solve and minimize through systems of ODEs, we can extend the previous method integrally defined method to multistate models which is the extension proposed by Groha, Schmon, and Gusev (2020).

As summarized earlier in fig. 3.1, the hazard is not the only quantity possible if one wants to completely characterize the survival distribution. We have seen that direct modelling of the survival S itself has proven to be highly successful: the NPMLE approach applied to the likelihood with the survival as the object of interest yields the Kaplan-Meier estimator as well as its conditional versions but it is also possible to directly estimate the density of event times. Estimation of p_Y can be achieved by multiple means: either by parametrization of a known parametric model, which is the approach taken by Ranganath et al. (2016) where a neural network parametrizes a deep exponential family such that

$$\begin{aligned} a &\sim \mathcal{N}(0, \sigma_a) \\ b &\sim \mathcal{N}(0, \sigma_b) \\ z_n &\sim \mathrm{DEF}(\theta) \\ x_n &\sim g(\cdot \mid \beta, z_n) \\ t_n &\sim \mathcal{W}\left(\log(1 + \exp(z_n^\intercal a + b)), k\right), \end{aligned}$$

where $\mathrm{DEF}(\theta)$ denotes a deep exponential family parametrized by θ, g is a prior distribution on the data, and $\mathcal{W}(\alpha, k)$ the Weibull distribution[84] with scale α and shape k. Similarly, in Martinsson (2016) a recurrent neural network is used to parametrize the parameters of a Weibull such that

$$T_i \sim \mathcal{W}\left(\mathrm{rnn}_1(X_1^i, \ldots, X_t^i), \mathrm{rnn}_2(X_1^i, \ldots, X_t^i)\right),$$

where X_t^i is the covariate vectors of individual i at time t. $(X_1^i, \ldots, X_t^i)$ is therefore the whole history of the covariates of for the individual i when those are time varying. While simpler than the deep exponential family approach, the previous approach is able to naturally deal with time-varying covariates and produce new predictions dynamically as new data is acquired. These approaches, while able to encode complex relations in the data by means of neural networks, are still constrained by the choice of parametric family. Attempts to model the density without such restrictive assumptions have been presented in C. Lee, Zame, et al. (2018); C. Lee, Yoon, and Schaar (2020) where they directly model the density without making any assumption on the distribution but do so by discretizing the space, an undesirable limitation. We have seen earlier that Groha, Schmon, and Gusev (2020) directly model λ_Y before solving an ODE to maximize the likelihood and nothing, however, prevents the same technique to be applied to the density p_Y directly provided one is careful to impose

84: See fig. 3.3.

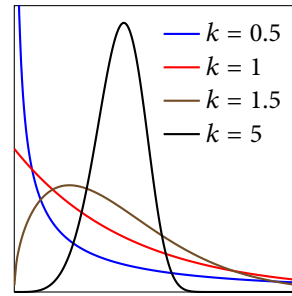


FIGURE 3.3: Weibull distribution for $\alpha = 1$ and varying k.

positivity as well as integration to 1. In practice those two constraints prove too hard to impose exactly and approaches based on penalizing towards these constraints have significant shortcomings (such as exploding logarithms) that makes then impractical.

In order to solve those significant shortcomings, and inspired by the approaches of Ranganath et al. (2016) and Martinsson (2016) where a simple variable is transformed to more closely resemble the survival distribution of interest, Miscouridou et al. (2018) propose the use of normalizing flows, that is mappings $m : \mathbb{R} \mapsto \mathbb{R}_+$ such that the survival Y is the image of a random variable $Z \sim \mathcal{Z}$ where $\mathcal{Z}$ is a simpler and known distribution i.e.

$$Y = m(Z).$$

The authors then show that a lower bound of the censored likelihood can be optimized, yielding a flexible estimator of the density of the survival distribution. This last approach, while used primarily by the authors as a way to build an estimator of the density is, however, strictly more powerful: by having access to the transformed distribution directly we are able to quickly build estimators of most of the quantities of interest as well as efficiently sample from the survival distribution by virtue of having access to a generative model. While sampling is often seen as a secondary concern,[85] it has enjoyed a renewed interest from the machine learning field. Given that the financial industry benefits greatly from access to efficient sampling methods in order to perform stress tests as well as complex Monte Carlo simulations, we are particularly interested in generative models that can also be used to compute accurate IPCW weights.

85: Sampling can often be performed later as a *by-product* by means of rejection sampling or other similar techniques.

3.3 Generative Models

Driven by the phenomenal advances in machine learning, the expectations of the possible applications from users and practitioners alike have risen in consequence. In fields such as image processing and natural language processing (NLP), while a decade ago the flagship problems were classification problems (and still are of great interest, as many seemingly more complex problems can be reframed as instances of classification problems), research and public interest has now shifted toward the problem of *generation*, that is the creation of new samples from a learned distribution.

Generative models have successfully been employed on images (see Brock, Donahue, and Simonyan [2019]), making it possible to generate realistic-looking images, potentially conditioned on some input. While from a purely theoretical point of view the task is interesting, one may still wonder "why on earth would I want to sample images?". These techniques have found a wide range of practical applications such as generating

FIGURE 3.4: These people do not exist.

FIGURE 3.5: Generating novel context specific animations without manual intervention.

character portraits in video games (Karras, Laine, Aittala, et al. [2020]), generating realistic character animations for tasks where no motion capture data is available (Holden et al. [2020]; Harvey et al. [2020]), transforming images conditionally on some inputs (Karras, Laine, and Aila [2019]), or augmenting data to create more robust training sets (Sandfort et al. [2019]). Similar techniques have been applied to speech, leading to tremendous improvements in text-to-speech (Yuxuan Wang et al. [2017]; Oord et al. [2016]) and opening up the possibility of artificially generating music (Dhariwal et al. [2020]). While less successful in their scope than the previous examples, attempts to use generative models with text have yielded many applications, from translation (Sutskever, Vinyals, and Le [2014]) and abstractive summarization (Lewis et al. [2020]), to automatic generation of financial reports from structured data (Rebuffel et al. [2020]) or augmentation of textual data (Dopierre, Gravier, and Logerais [2021]).

3.3.1 Generative Models in Finance

While usually not studied from the perspective of machine learning ,[86] the majority of models developed for the purpose of financial modelling[87] have historically been generative models. The models used to price derivatives usually involve the use of parametric generative models describing the

86: Or even statistics.

87: In the sense of modelling the financial objects tradable on the markets.

distribution of the underlying object, such as the Heston model (Heston [1993]) described by the stochastic differential equation (SDE)

$$\begin{aligned} \mathrm{d}S_t &= \mu S_t \,\mathrm{d}t + \sqrt{\nu_t} S_t \,\mathrm{d}W_t \\ \mathrm{d}\nu_t &= \theta(\omega - \nu_t)\,\mathrm{d}t + \xi\sqrt{\nu_t}\,\mathrm{d}B_t \\ \mathrm{d}W_t\,\mathrm{d}B_t &= \rho\,\mathrm{d}t, \end{aligned}$$

where W_t and B_t are standard Wiener processes; or the "stochastic alpha, beta, rho" (SABR) model (Hagan, Lesniewski, and Woodward [2015]) defined by

$$\begin{aligned} \mathrm{d}S_t &= \sigma_t S_t^{\beta}\,\mathrm{d}W_t \\ \mathrm{d}\nu_t &= \alpha\nu_t\,\mathrm{d}B_t \\ \mathrm{d}W_t\,\mathrm{d}B_t &= \rho\,\mathrm{d}t, \end{aligned}$$

which are then calibrated using the prices of derivatives observed on the market.[88] Once those generative models have been calibrated, they can be used to generate synthetic data, for example to test trading strategies, to perform backtests or even for anomaly detection. While those methods have been primarily developed because closed-form solutions exist and made cheap computations possible, recent trends have been toward more expensive and less interpretable models in exchange of more flexibility. Henry-Labordere (2019) proposes to learn nonparametric generative models as an alternative to the SABR model for the generation of financial data as well as anomaly detection. Similarly, Takahashi, Y. Chen, and Tanaka-Ishii (2019); Binkowski, Marti, and Donnat (2018) use generative models for time series and Marti, Goubet, and Nielsen (2021); Marti (2020) use generative adversarial network (GAN) to sample correlation matrices for portfolio stress testing. The Basel III as well as the future Basel IV regulations (Basel Committee [2018]) have made stress testing of credit risk mandatory but give financial actors some liberty on the choice of credit model as well as stress test methodology.[89] In order to not only fulfil its regulatory duties but also minimize the safety cushion it needs to maintain, a financial actor is strongly incentivised toward highly accurate models for the purpose of stress testing. As credit decisions are usually modelled as complex hierarchical processes involving defaults, contagions and loss recovery; it is easy to see how conditional generative models of the different sub-components can be useful and have a sizeable financial impact. Given their usefulness as well as our need for estimators of the survival, we will dedicate the rest of the chapter to a specific generative model. While many generative models do indeed match our requirements of easy sampling and tractable survival as well as likelihood, by virtue of specifying parametric families of known distributions, the only requirement usually required is the ease of sampling. GANs for example, have

88: While out of the scope of this thesis, we quickly explain how traditional mathematical finance works as it happens to use very similar techniques than those used in survival analysis and recently in diffusion flows. Usually an asset is assumed to follow some diffusion process governed by an SDE then, under some assumptions, the price of a derivative of that asset can then be assumed to be equal to the expected value of some functional of the diffusion under a risk neutral martingale measure. The price of the derivative can then be computed by writing a Kolmogorov or Fokker-Planck equation and solving the corresponding partial differential equation (PDE).

89: Within reason; any non-standard methodology needs to be motivated before the regulator.

FIGURE 3.6: Mapping a normal distribution to a survival distribution.

enjoyed a growing popularity for the quality of the samples they provide for complex distributions but even accessing the likelihood is intractable. Even the more standard models such as VAEs (Kingma and Welling [2014, 2019]), are usually constrained to lower bounds of the likelihood and are therefore unsuitable for our use. Given those prerequisites, we introduce in the next subsection a specific recent spiritual successor of VAE with the added benefit of a tractable likelihood.

3.3.2 Normalizing Flows

Similarly to regression models such as the AFT model of eq. (3.8), we can assume that $Y \sim \Upsilon$ can be obtained as a transformation through some mapping m of some latent variable $Z \in \mathbb{Z}$ with $Z \sim \mathcal{Z}$, as represented in fig. 3.6, such that

$$Y = m(Z, X). \tag{3.9}$$

This view is not restrictive as, when all the variables are continuous, we can indeed always write,

$$Y = F^{-1}_{Y|X=x} \circ F_Z(Z),$$

knowing that $F_Z(Z) \sim \mathcal{U}[0,1]$ and $F^{-1}_{Y|X=x}(U) \sim \Upsilon$ when $U \sim \mathcal{U}[0,1]$. As $F_{Y|X=x}$ is unknown, it is natural to instead select the best candidate $m_{\theta^\star}$ from some family $(m_\theta)_\theta$, parameterized by $\theta \in \Theta$ (usually $\Theta = \mathbb{R}^p$), that minimizes some notion of distance to the true distribution Υ i.e. such that

$$g_{\theta^\star} = \underset{\theta \in \Theta}{\operatorname{argmin}}\, \mathcal{D}(\mu_\Upsilon, m_\theta(\mu_\mathcal{Z}, X)), \tag{3.10}$$

where $m_\theta(\mu_\mathcal{Z}, X)$ denotes the push-forward measure of $\mu_\mathcal{Z}$ by the mapping $m_\theta(\cdot, X)$. Several distances $\mathcal{D}$ have been proposed in the censored setting, such as the partial log-likelihood introduced earlier or continuous ranked probability score (CRPS) (Avati et al. [2018]) for sharp estimates but we will only consider the former.

The task of learning a generative model of Y can then be seen from a different optic: instead of trying to learn a complex distribution we learn

a complex mapping of a simple distribution. We therefore need to be able to solve eq. (3.10), which can be done if the problem, and therefore both $\mathcal{D}$ and m_θ, have a particular form amenable to a simple solution[90] or if both $\mathcal{D}$ and m_θ are differentiable.[91]

In the following sections, we will see how to construct the mapping of eq. (3.9) such that the mapping m_θ itself as well as the quantities involved in eq. (3.2) are both computable and differentiable. While differentiability alone is enough, we require in practice that the different objects involved can be *automatically differentiated*, i.e. that it is possible to algorithmically derive the *exact* gradient as opposed for example to finite differences. *Automatic differentiation* of some function f can refer either to *forward mode* automatic differentiation (AD) where the quantity being algorithmically computed is $J_f(x)v$ where $J_f(x)$ is the jacobian of f evaluated at x and v is some vector. $J_f(x)v$ is then simply the directional derivative and given a set of rules to build vector-jacobian product Jv, it is easy to see that the rule for $f \circ g$ can be obtained by $J_g(J_f(x)v)$.[92] *Reverse mode* AD on the other hand is equivalent to the calculation of $J_f^\intercal(x)v$,[93] which results in a reversed rule for composition i.e. $v^\intercal J_{f\circ g}(x) = (v^\intercal J_g(x))J_f(x)$ which algorithmically implies that the process of deriving the full vector-jacobian product proceeds in two steps: in the *forward* pass, the composition is executed and recorded, for example as a tape or graph of operations, and once the output obtained the tape is played backward in the *reverse* phase in order to compose the elementary vector-jacobian products. If $f \circ g(x) \in \mathbb{R}$ then taking $v = 1$ yields $v^\intercal J_{f\circ g}(x) = \nabla(f \circ g)$, which motivates the use of reverse mode AD in machine learning[94] where cheap computation of the gradient makes gradient descent computationally viable. In practice, for a function $f : \mathbb{R}^p \mapsto \mathbb{R}^m$, forward differentiation scales linearly with p, that is has a total computation complexity of $O(p)$ while reverse mode scales with $O(m)$. As most machine learning problems involve a significant number of parameters (p large) and small numbers of outputs (usually $m = 1$), reverse mode AD is the umost commonly used mode in the machine learning setting. This is however not always the case and in some instances, such as the computation of a Hessian, mixed mode AD[95] can be employed to exploit the specific tradeoffs of the problem at hand. For an overview of the field of *automatic differentiation* in general we refer to Margossian (2019). Given the recent successes of the field of deep-learning, the desire to replace usually simple or even fixed functions with complex neural networks have given rise to the need to formulate problems such that the whole end-to-end pipeline is differentiable. This new field, usually refered to as *differentiable programming* or *scientific machine learning* (see Innes et al. [2019]), relies on building adjoints, that is building rules for the construction of the vector-jacobian product $v \mapsto v^\intercal J(x)$ in order to enable the use of increasingly complex or exotic "layers" inside end-to-end machine

90: For example the task of minimizing the Kullback–Leibler (KL) divergence or log-likelihood from a shifted and scaled Weibull.

91: Of course differentiability is not sufficient to ensure that minimization is possible, convexity would normally be required, but local minimums are in practice considered *good enough*.

92: Forward mode is usually implemented using dual numbers, i.e. by operating on $\mathbb{R} + \epsilon\mathbb{R}$ where $\epsilon^2 = 0$ instead of $\mathbb{R}$.

93: Or $v^\intercal J_f(x)$.

94: Usually under the name *backpropagation*.

95: *Forward-over-reverse* in this case

learning systems. Recently, methods to make convex optimization[96] differentiable have been proposed (Agrawal et al. [2019]), as well as approaches to differentiable sorting[97] (Cuturi, Teboul, and Vert [2019]; Blondel et al. [2020]), or differentiable solvers for ODEs, PDEs, SDEs (Rackauckas and Nie [2017a]) and universal differential equation (UDE) (Rackauckas, Ma, Martensen, et al. [2020]) in general. We will, in particular, make use of these last developments in the sections that follow.

96: Not the *solution* itself, but the *program* that takes the parameters of the problem as input and give the solution as output.

97: While we will not use it here, differentiable sorting enables differentiable computations of the AUC or C-Index, both useful metrics in the censored setting.

3.3.3 The Change of Variable Theorem

We momentarily only concern ourselves with the *unconditional* estimation of the density and associated survival and omit the conditioning on the covariates X. This is only for the sake of readability and one needs only to introduce the missing conditioning in all the following relations to retrieve the conditional case.

Assuming the existence of the mapping $Y = m_\theta(Z)$ and under the hypothesis that m_θ is a C^1-diffeomorphism it is possible to derive the density of Y from the density of Z by means of the change of variable theorem

$$\log p_Y(t) = \log p_Z(z) - \log\left|\det \frac{\partial m_\theta}{\partial z}\right|. \tag{3.11}$$

Equation (3.11) imposes not only the explicit constraint that m_θ must be invertible[98] but also that the determinant of the Jacobian is easy to compute. Such constraints are in practice fairly difficult to meet and impose restrictions on the parametric family (m_θ) in order to render the problem tractable. In the general setting, computing the determinant of an arbitrary matrix of size $q \times q$ has a computational cost[99] of $O(q^3)$. In practice we restrict m_θ such that its Jacobian has a form facilitating the computation of its determinant i.e., for example triangular or block diagonal. It is, however, possible to retrieve the lost representational power by simply composing multiple simple transformations such that

$$\begin{aligned} Y &= m_{\theta,K} \circ \ldots \circ m_{\theta,0}(Z), \\ \log p_Y(t) &= \log p_Z(z) - \sum_{i=0}^{K} \log\left|\det \frac{\partial m_{\theta,i}}{\partial z_i}\right|. \end{aligned} \tag{3.12}$$

Since the original paper of Rezende and Mohamed (2015) most of the research on the subject has gravitated toward constructing families of functions m_θ that are easily computable and whose Jacobian has a tractable determinant i.e. diagonal, triangular, has a simple block structure or more generally encodes an adjacency matrix (Wehenkel and Louppe [2021]) while still maintaining a high degree of flexibility. Dinh, Krueger, and

98: Hence the C^1-diffeomorphism assumption.

99: Computing a determinant can be stated in terms of matrix multiplication and has the same complexity. This therefore means that the best known algorithm has a complexity of $O(q^{2.3729})$(Alman and Williams [2020]) but with a hidden constant so large it is impractical. For all intents and purposes, the fastest algorithm in practice is $O(q^{2.807})$(Strassen [1969]), with both lower-upper decomposition (LU) and Bareiss often used leading to the aforementioned $O(q^3)$(Bareiss [1968]).

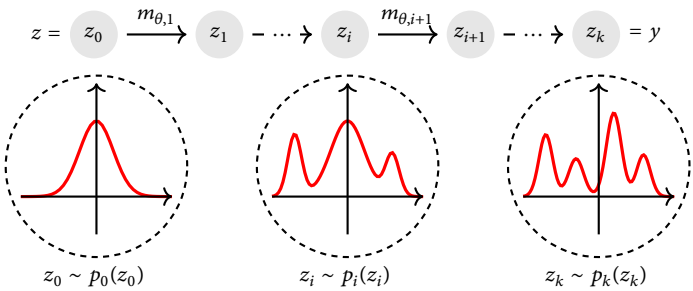


FIGURE 3.7: Mapping a simple distribution to a target distribution by successive compositions.

Y. Bengio (2015) introduced the notion of *coupling layers*, that is layers that leave part of the input untouched in order to ensure invertibility:

$$
\begin{aligned}
m_\theta(x)_{1:k} &= z_{1:k}, \\
m_\theta(x)_{1:k} &= z_{k+1:d} \odot \exp(s(z_{1:k})) + t(z_{1:k}),
\end{aligned}
$$

where s and t are respectively scaling and translation operations from $\mathbb{R}^k \mapsto \mathbb{R}^{d-1}$. Not only is it possible to ensure invertibility but the resulting Jacobian is block triangular

$$
J = \begin{pmatrix} \mathbb{I}_k & \mathbb{0}_k \\ \frac{\partial y_{k+1:d}}{\partial z_{1:k}^\intercal} & \operatorname{diag}(\exp(s\,(z_{1:k}))) \end{pmatrix}.
$$

In order to be able to compose those coupling layers, Dinh, Sohl-Dickstein, and S. Bengio (2017) subsequently proposed the use of masked convolutions in order to vary the split between the untouched and transformed part of the coupling. Similarly, Oliva et al. (2018); Papamakarios, Pavlakou, and Murray (2018) show that autoregressive networks exhibit the same structure as the aforementioned coupling layers and can therefore be used as candidates in normalizing flows. Kingma and Dhariwal (2018) builds on the previous approaches and introduce invertible 1×1 layers, further improving the representational power of the individual flows.

It is, however, possible to entirely sidestep the previous problems by defining eq. (3.12) continuously, as proposed by R. T. Q. Chen et al. (2018). By interpreting the change of variable associated to a single composition step as an Euler integration step from i to $i+1$,[100] we can instead adopt an infinitesimal point of view by parametrizing the derivative of the change of variable. It is possible to prove (see the derivation in R. T. Q. Chen et al. [2018], Appendix A) that the change of variable theorem becomes:

$$
\begin{cases} z_{i+1} = m_{\theta,i}(z_i) \\ \dfrac{\log p(z_{i+1}) - \log p(z_i)}{i+1-i} = -\log\left|\dfrac{\partial m_{\theta,i}}{\partial z}\right| \end{cases} \Rightarrow \begin{cases} \dfrac{\partial \mathbf{z}_\theta}{\partial t} = m_\theta(\mathbf{z}_\theta(t), t) \\ \dfrac{\partial \log p(\mathbf{z}_\theta(t))}{\partial t} = -\operatorname{tr}\dfrac{\partial m_\theta}{\partial \mathbf{z}} \end{cases}.
$$

100: One can interpret $m_{\theta,i}(z_i)$ as $m_\theta(z_i, i)$.

We distinguish $\mathbf{z}(t)$ the path of the flow, from z the initial latent variable. Here $\mathbf{z}(t)$ is only a mathematical device used to define the transformation of interest and is not in itself the object of interest.

The compositional process of eq. (3.12) is therefore replaced by the initial value problem of eq. (3.13)

$$\begin{aligned} \frac{\partial}{\partial t}\begin{bmatrix} \mathbf{z}_\theta(t) \\ \log p(y) - \log p(\mathbf{z}_\theta(t)) \end{bmatrix} &= \begin{bmatrix} m_\theta(\mathbf{z}_\theta(t), t) \\ -\operatorname{tr} \dfrac{\partial m_\theta}{\partial \mathbf{z}} \end{bmatrix} \\ \begin{bmatrix} \mathbf{z}_\theta(1) \\ \log p(y) - \log p(\mathbf{z}_\theta(1)) \end{bmatrix} &= \begin{bmatrix} y \\ 0 \end{bmatrix}. \end{aligned} \tag{3.13}$$

Note that the problem as written here defines the flow in the direction $\mathcal{Y} \mapsto \mathcal{Z}$ i.e. the mapping from Y to Z. The inverse mapping $\mathcal{Z} \mapsto \mathcal{Y}$ is similarly defined by changing the starting point and matching initial conditions of the problem. Note that for fixed endpoints Y and Z, the path stays the same and only the direction of the dynamics is reversed, which is equivalent to reversing the time. We denote by $\mathbf{z}_\theta(t, z_1)$ (resp. $\mathbf{z}_\theta(t, z_0)$) the path solution of the particular instance of the initial value problem (IVP) of eq. (3.13) with initial condition $\mathbf{z}_\theta(1, z_1) = z_1$ (resp. $\mathbf{z}_\theta(0, z_0) = z_0$) and parametrization θ of the dynamics. By parameterizing a normalizing flow infinitesimally we are able to overcome the two previous limitations: not only the expensive computation $O(n^3)$ of the determinant is entirely eliminated and replaced by a trace operation costing only $O(n)$ but the restriction on invertibility is not explicitly required anymore: as long as m_θ and $\partial_{\mathbf{z}} m_\theta$ are piecewise continuous and Lipschitz over the integration domain, then eq. (3.13) admits a unique solution (see Khalil [2002], Theorem 3.2, p.93; or Hirsch, Smale, and Devaney [2013]), which by construction is the mapping desired and therefore a C^1-diffeomorphism and inverting the solution only requires solving the ODE backward in time. In practice these hypotheses are met for most of the common layers and activation functions used in deep learning (see Scaman and Virmaux [2018]).

We note here that while m_θ in the continuous definition of the normalizing flows plays a similar role to m_θ in the discrete version, it is in fact not the same object and does not represent the flow explicitly but only implicitly through the dynamics. Instead, for $\mathbf{z}_\theta(\cdot, z_0)$ solution of eq. (3.13) with initial value $\mathbf{z}_\theta(0) = z_0$, we denote by M_θ the resulting normalizing flow such that

$$\begin{aligned} M_\theta(z) &\overset{\text{def}}{=} \mathbf{z}_\theta(1, z_0), \\ M_\theta^{-1}(z) &\overset{\text{def}}{=} \mathbf{z}_\theta(0, z_1). \end{aligned} \tag{3.14}$$

It is not possible[101] to ensure that the learned flow maps from $\mathbb{Z}$ to $\mathbb{R}_+$ exactly, where $\mathbb{Z}$ is the support of the pullback distribution $\mathcal{Z}$.[102] While,

101: Or at least not *trivially* possible.

102: Considered an hyperparameter to be chosen by the user.

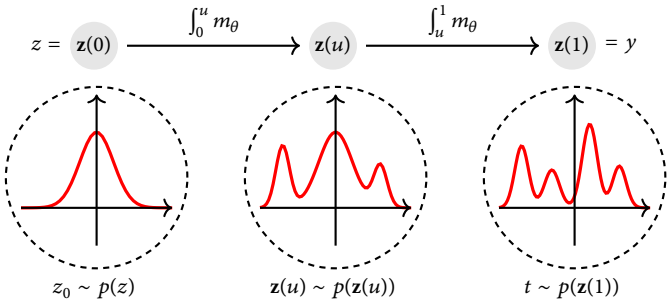


FIGURE 3.8: Mapping a simple distribution to a target distribution by continuously applying an infinitesimal flow.

barring any numerical inaccuracies, this is indeed true by construction for any $(Z_i, M_\theta(Z_i))$ part of the training sample, we cannot guarantee that the same is true for new observations. Indeed there is no reason for the learned mapping to map to a proper survival distribution, that is for the pushforward measure to have support in $\mathbb{R}_+$. In order to bypass any discussion and entirely eliminate the problem, we instead reparametrize the normalizing flow using a two-step process. First a learned flow maps from $\mathbb{Z}$ to $\mathbb{R}$ by learning

$$\log(Y) = M_\theta(Z),$$

then a deterministic change of variable maps $\mathbb{R}$ to $\mathbb{R}_+$ through

$$Y = \exp\big(\log(Y)\big) = \exp(M_\theta(Z)).$$

The resulting process is still a normalizing flow[103] and could be reframed as a continuous normalizing flow by deriving the corresponding dynamics. This reparameterization ensures that $M_\theta(Z) \in \mathbb{R}_+$ is a proper time-to-event as well as prevents any issues arising from the possibility of mapping an event outside the support of the latent distribution. This reparameterization is commonly used in survival *regression* (Miller and Halpern [1982]; Buckley and James [1979]) and our model can be seen as a generalization of the accelerated failure time model of eq. (3.8) (see Wei [1992]). For simplicity, however, we perform the last change of variable independently and we omit this last step from the notation in the following sections for the sake of readability.

103: By seeing exp as a *layer*, which is indeed a C^1-diffeomorphism.

3.4 Unconditional Survival Normalizing Flows

The observant reader may have already noticed earlier that the principal argument we put forward for the use of *continuous* normalizing flows, that is the fact that computation of the Jacobian is trivial, is not convincing or

even necessary in the specific case of survival analysis. Because normalizing flows have to be bijective, unlike VAEs, it is necessary to preserve the dimensions of the input and output spaces. As a consequence, flows in our setting are from $\mathbb{R}$ to $\mathbb{R}$ only and the Jacobian of the flow is a degenerate matrix of size 1×1, whose determinant is not hard to compute. We therefore adopt here the continuous approach mainly for the two other reasons evoked earlier: the absence of restrictions on the choice of m_θ, which while not particularly problematic simplifies experimentations, and more importantly the ability to compute the inverse flow mapping exactly and for the same computational price as the mapping flow itself. This last characteristic will prove to be particularly important in the survival setting compared to the usual density estimation setting as we will detail here.

While the system given previously in eq. (3.13) describes how to obtain $Y_i = M_\theta(Z_i)$ as well as $f_T(Y_i) = \log p(M_\theta(Z_i))$, we also need to be able to compute $S_Y(Y_i)$ in order to entirely define the loss, chosen here to be the partial likelihood of eq. (3.2). We exploit the relation between Y and Z and note that,

$$\begin{aligned} S_Y(Y_i) &= \int_{\mathbb{R}_+} \mathbb{1}_{t \geq Y_i} \, \mu_Y(\mathrm{d}t) = \int_{\mathbb{R}_+} \mathbb{1}_{M_\theta(z) \geq Y_i} \, M_\theta\left(\mu_Y(\mathrm{d}z)\right) \\ &= \int_{\mathbb{Z}} \mathbb{1}_{z \geq M_\theta^{-1}(Y_i)} \, \mu_Y(\mathrm{d}z) \\ &= S_Z\left(M_\theta^{-1}(Y_i)\right), \end{aligned} \tag{3.15}$$

where the penultimate equality is not trivial but can be obtained by taking the derivative of $M_\theta(z) = z(1, z, \theta)$ with respect to z which yields, provided m_θ is sufficiently smooth, an adjoint initial value problem whose dynamical system is loosely decoupled, with adjoint state

$$\frac{\mathrm{d}}{\mathrm{d}z} M_\theta(z) = \exp\left(\int_0^1 -\frac{\partial}{\partial z} m_\theta\left(\mathbf{z}_\theta(t, z), t\right) \mathrm{d}t \right) > 0.$$

We do not solve the adjoint system and only use the fact that it is positive in order to determine the sense of the inequality. From eq. (3.15), we see that computing the survival function on the unknown distribution $\mathcal{Y}$ can be reframed as computing the survival of the preimage on the, supposed entirely known, pullback distribution $\mathcal{Z}$. By construction, computing $M_\theta^{-1}(Y_i)$ is of the same complexity as computing $M_\theta(Z_i)$ and only requires solving the system backward in time. Now that we have shown the quantities to be *computable*, we explain how to efficiently obtain the gradient of these quantities.

3.4.1 Parameter Estimation

We recall that we wish to find θ, solution of the empirical risk minimization problem of eq. (3.2). As the use of computationally expensive neural networks constrains our choice of optimization techniques to first order descent methods only, we only need to be able to differentiate M_θ, solution of the IVP of eq. (3.13).

As all the quantities involved are differentiable, we only need to be able to compute the sensitivity $\partial_\theta M_\theta(Z)$ of the solution of the ODE itself with respect to its parameters, usually referred to in the ODE literature as local sensitivity analysis (as opposed to global sensitivity analysis which concerns itself with the study of the range of solutions given the whole feasible domain of inputs and parameters). By rewriting $M_\theta(Z)$ as

$$\begin{aligned} M_\theta(Z) &= \mathbf{z}_\theta(1, \theta) \\ &= \int_{t_0}^{T} \mathbf{z}_\theta(t, \theta)^\intercal \begin{bmatrix} 0 \\ 1 \end{bmatrix} \delta_1(t) \, \mathrm{d}t, \end{aligned}$$

the loss of eq. (3.2), given a solution $\mathbf{u}$ of the IVP of eq. (3.13) (with $\mathbf{u} = [\mathbf{z}_\theta, \Delta \log p(\mathbf{z}_\theta)]$), can be written as

$$L(\mathbf{u}, \theta) = \int_{t_0}^{T} l(\mathbf{u}(t, \theta), \theta) \, \mathrm{d}t. \tag{3.16}$$

We can then form the adjoint state

$$\begin{aligned} \frac{\partial \lambda}{\partial t} &= \frac{\partial l}{\partial u}(\mathbf{u}(t, \theta), \theta) - \lambda(t) \frac{\partial p}{\partial u}(t, \mathbf{u}(t, \theta), \theta), \\ \lambda(T) &= 0, \end{aligned}$$

such that

$$\frac{\partial L}{\partial \theta} = \int_{t_0}^{T} \lambda(t) \frac{\partial p}{\partial \theta}(\mathbf{u}(t, \theta), \theta) + \frac{\partial l}{\partial \theta}(\mathbf{u}(t, \theta), \theta) \, \mathrm{d}t + \lambda(t_0) \frac{\partial \mathbf{u}}{\partial \theta}(t_0, \theta).$$

This adjoint method (see Cao et al. [2003]) doesn't require any additional machinery (other than a $O(1)$ increase in computations) in order to obtain the derivative and the solution itself, we only need to solve the original IVP with the addition of the new adjoint state, resulting in the same asymptotic computational complexity.[104]

Using the adjoint method, we are therefore able to differentiate $M_\theta = \mathbf{z}_\theta(1, \cdot)$ as well as $M_\theta^{-1} = \mathbf{z}_\theta(0, \cdot)$ with respect to θ. We refer to Rackauckas, Ma, Dixit, et al. (2018) for a more complete overview of the various possibilities for automatic differentiation of a solution of a differential equation as well as Innes et al. (2019) for the perspectives offered.

104: And a constant increase in practice.

3.5 Conditional Survival Normalizing Flows

In the previous sections, we omitted the conditioning on the covariates $X \in \mathbb{X}$ in order to simplify the notations. We now reintroduce this conditioning, as its presence does not modify any of the previous results, and show how to efficiently build a mapping $M_\theta : \mathcal{Z} \mapsto \mathcal{Y} \mid X$ such that

$$Y = M_\theta(Z, X).$$

As the trace operator is linear, it is possible to efficiently extend the expressivity of a single normalizing flow at a minor computational cost by representing it as a linear combination of K basis functions i.e.

$$m_\theta(\mathbf{z}_\theta(t, x), t, x) = \sum_{i=1}^{K} m_{\theta,i}(\mathbf{z}_\theta(t, x), t, x).$$

As in R. T. Q. Chen et al. (2018), we choose to parametrize each basis function $m_{\theta,i}$ as a mixture of unconditional and time-invariant dynamics. We choose to decouple the gating in x and t in order to prevent overfitting and be able to apply different regularizations and computational budget. Decoupling the blocks in x, t and z also gives us the ability to exploit the structure of the problem in order to implement efficient batching for use on graphical processing units (GPUs).

$$m_\theta(\mathbf{z}_\theta(t, x), t, x) = \sum_i \pi_{\theta,i}(x)\sigma_{\theta,i}(t)m_{\theta,i}(\mathbf{z}_\theta(t, x)).$$

The full dynamics of our continuous normalizing flow are therefore,

$$\begin{aligned} \frac{\partial}{\partial t}\mathbf{z}_\theta(t, x) &= \sum_i \pi_{\theta,i}(x)\sigma_{\theta,i}(t)m_{\theta,i}(\mathbf{z}_\theta(t, x)), \\ \frac{\partial}{\partial t}\log p(\mathbf{z}_\theta(t) \mid x) &= -\sum_i \pi_{\theta,i}(x)\sigma_{\theta,i}(t)\,\mathrm{tr}\,\frac{\partial m_{\theta,i}}{\partial \mathbf{z}}\Big|_{\mathbf{z}_\theta(t,x)}. \end{aligned} \tag{3.17}$$

3.5.1 Hierarchical Conditioning

Conditional density estimators learned by maximizing the likelihood are well known to be prone to overfitting. We control the amount of overfitting by introducing an auxiliary latent representation shared by all the conditional distributions $Y \mid X$. We therefore impose the shared hierarchical representation $w = H_\theta(z)$ such that $y = M_\theta(w, x) = M_\theta(H_\theta(z), x)$ with $H_\theta(Z) \perp\!\!\!\perp X$. The corresponding flow dynamics can be rewritten as

$$\frac{\partial \mathbf{z}_\theta}{\partial t}(t, x) = \sum_{i=1}^{K}\Big(\mathbb{1}_{t\leq t_x} + \pi_{\theta,i}(x)\mathbb{1}_{t>t_x}\Big)\,\sigma_{\theta,i}(t)m_{\theta,i}(\mathbf{z}_\theta(t, x)), \tag{3.18}$$

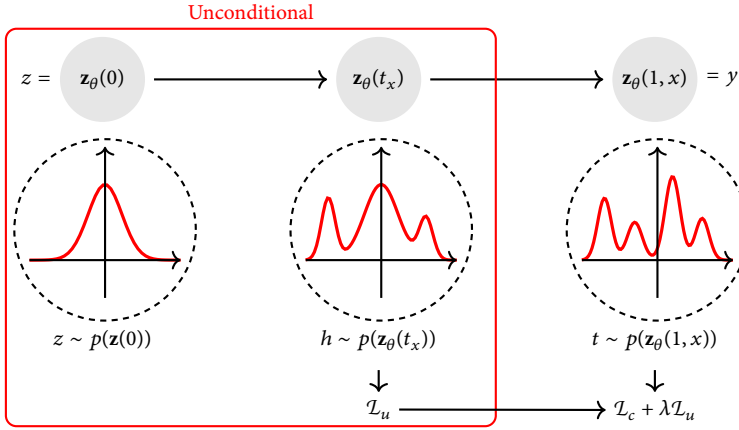


FIGURE 3.9: Hierarchical Survival Flow with Multiple Losses.

where t_x controls implicitly the allowed deviation from the unconditional distribution. If we denote by $p_{H_\theta(z)}$ the unconditional distribution induced by $H_\theta(z)$ and $p_{m_\theta(H_\theta(z),x)}$ the distribution induced by $p_{m_\theta(H_\theta(z),x)}$, we can regularize the intermediate shared latent representation toward the true *unconditional survival distribution* by augmenting the loss eq. (3.2) with an intermediary loss $\mathcal{L}_u$ such that

$$\mathcal{L}_{\text{total}}(\theta) = \mathcal{L}_u(\theta) + \mathcal{L}_c(\theta) \tag{3.19}$$

$$= \sum_i \delta_i \log(p_{H_\theta(Z_i)}(T_i)) + (1 - \delta_i) \log(S_{H_\theta(Z_i)}(T_i)) \tag{3.20}$$

$$+ \lambda \sum_i \delta_i \log(m_\theta(H_\theta(Z_i), X_i)(T_i)) + (1 - \delta_i) \log(S_{m_\theta(H_\theta(Z_i),X_i)}(T_i)),$$

where $S_{H_\theta(x)}$ and $S_{m_\theta(H_\theta(z),x)}$ are the respective survival functions of the distributions defined by the densities $p_{H_\theta(z)}$ and $p_{m_\theta(H_\theta(z),x)}$.

3.5.2 Discrete & Continuous Hierarchical Conditioning

For datasets with particularly complex dependences on the covariates, it is possible to add other layers of hierarchies. The simplest scheme consists of using *discrete* hierarchical transformations: let H be the number of hierarchies then we can learn $K \times H$ mixtures such that

$$m_\theta(\mathbf{z}_\theta(t, x), t, x) = \sum_{h=1}^{K} \mathbb{1}_{t_{h-1} < t \leq t_h} \sum_{i=1}^{K} \pi_{\theta,i,h}(x) \sigma_{\theta,i,h}(t) m_{\theta,i,h}(\mathbf{z}_\theta(t, x)). \tag{3.21}$$

We call this scheme *discrete hierarchical conditioning*. In the same manner as done in the previous section, it is possible if desired to introduce intermediary losses in order to prevent overfitting as well as help with the training procedure. As our model is continuous, we do not have to constrain ourselves to hard gating at time steps (t_h), effectively reducing our model to a standard normalizing flow using continuous normalizing flow layers. Instead we can *continuously* interpolate between representations:

$$m_\theta(\mathbf{z}_\theta(t,x),t,x) = \sum_{h=1}^{H} \exp\left(c_h|t-t_h|^2\right) \sum_{i=1}^{K} \pi_{\theta,i,h}(x)\sigma_{\theta,i,h}(t) m_{\theta,i,h}\left(\mathbf{z}_\theta\left(t,x\right)\right). \quad (3.22)$$

We call this scheme *continuous hierarchical conditioning*. By making c_h and t_h learnable parameters, we make it possible for the model to learn if hierarchies are needed or can potentially be discarded.

The hierarchical approach as introduced previously may at first appears strictly identical to the non-hierarchical approach as it is possible to incorporate $\mathbb{1}_{t_{h-1}<t\leq t_h}$ directly inside σ, there is, however, one significant advantage: by knowing the relative order (in the sense of t) in which the mixtures are applied we can design $m_{\theta,i,h}$ to be of increasing complexity in order to impose a shared representation between the different individuals.

3.6 Experiments

All the material, including code and data, necessary for the reproduction of the results presented here is available at `git.sr.ht/~aussetg/nfsurvival`. The experiments have been implemented in the Julia language (Bezanson et al. [2017]). where we make heavy use of the `DifferentialEquations.jl` (Rackauckas and Nie [2017b]) and `Zygote.jl` (Innes et al. [2019]) libraries in order to implement automatic differentiation of the initial value problems involved. The normalizing flow approach is directly compared to existing methods for survival analysis on synthetic data designed to model violations of the proportional hazards hypothesis as well as multimodality. We compare our approach to the existing literature on standard open medical datasets and motivate the need for generative models that are easy to sample from by applying our method to a commonly encountered setting in the financial community later in chapter 5.

3.6.1 Synthetic Data

In order to test the ability to both capture complex interactions between covariates as well as model a potentially multimodal distribution which

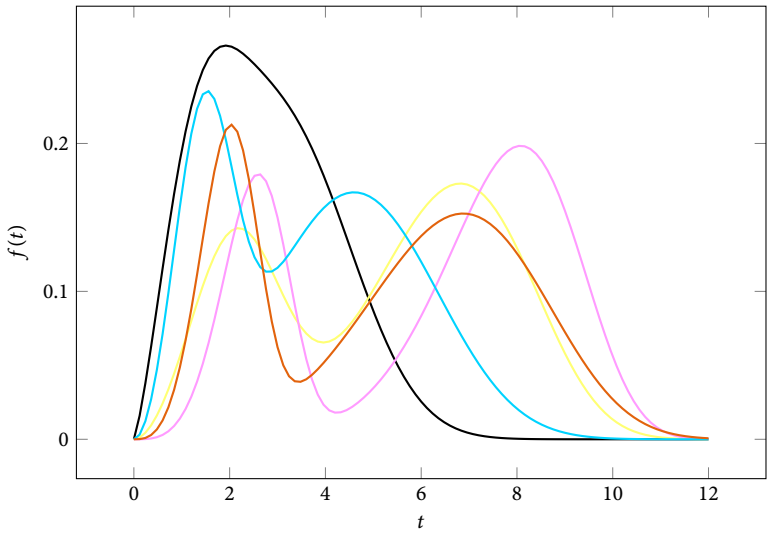


FIGURE 3.10: Different Synthetic Distributions for $d = 10$

violates the proportional hazard assumption, we generate synthetic data according to the following model:

$$
\begin{aligned}
X &\sim \mathcal{U}_d \\
Y &\sim pW(\beta_1^\intercal X, \beta_2^\intercal X) + 0.7W(2\beta_3^\intercal X, \beta_4^\intercal X) \\
C &\sim W(\beta_5^\intercal X, \beta_6^\intercal X) \\
T &= \min(Y, C) \\
\delta &= \mathbb{1}_{Y \leq C},
\end{aligned}
$$

with some resulting distributions represented in §3.6.1. While often implicitly discarded by the model chosen, the possibility for the event distribution to be multimodal is far from exotic: many diseases, such as acute radiation poisoning, include a latent period of relative well-being of the patients; death occurring before or after the latent period but not during. Similarly, in a financial setting it is expected to observe modes around important fiscal events as those are the periods around which one expects a company to default. The different models are trained on a test set of 3000 observations (chosen to match the characteristics of the real datasets) with an average of 80% of censoring. We then estimate the Harrel's concordance index (Harrell, Califf, et al. [1982]; Harrell, K. L. Lee, and Mark [1996])

$$
\mathbb{P}_Y \left(s(X_i) \geq s(X_j) \mid Y_i \leq Y_j \right), \tag{3.23}
$$

where $s(X_i)$ is a scoring function defined later and (X_i, Y_i) and (X_j, Y_j) are identically distributed. The concordance index has been widely used in

the survival setting as it can be corrected to account for censoring and only concerns itself with the global *ranking* capabilities of the model instead of the accuracy of the predictions. While incomplete and criticizable, the concordance metric represents one important aspect of the model for practitioners: how accurate are the relative risks (see Uno et al. [2011]). This is particularly important in the medical or financial setting where whether to select (for treatment or for financing) an individual over another is the useful actionable insight.

While proportional hazard methods possess a natural notion of risk score, needed to compute the concordance index, this is not the case for our method or survival forests. In the random survival forest setting (Ishwaran, Kogalur, et al. [2008]) the authors construct the risk score as

$$s(X_i) = \sum_{i=1}^{m} \hat{H}_i(t_i \mid X_i), \tag{3.24}$$

where $(t_i)_{i=1,\dots,m}$ are the unique event times in the training set. While a similar approach can be used for our model, we instead exploit the fact that we can easily and cheaply generate conditional observations to directly learn the ranking implied by the concordance. Given a test dataset $(T_i, \delta_i, X_i)_{i=1,\dots,n}$ we first generate $n \times m$ observations $Y_{i,k} = G_\theta(Z_k, X_i)$ with $Z_k \sim \mathcal{Z}$ i.i.d. and define the score vector $\mathbf{s} = [s(X_1), \dots, s(X_n)]$ as

$$s(X_i) = \frac{1}{m} \sum_j \sum_k \mathbb{1}_{Y_{i,k} > Y_{j,k}}. \tag{3.25}$$

The competing approaches have been trained using the `PySurvival` (Stephane [2019]) library and their concordance computed using the scoring function from the same package. The results are presented in table 3.1.

Method	Concordance
This work	**0.795691**
DeepSurv (Katzman et al. [2018])	0.762831
Random Survival Forest (Ishwaran, Kogalur, et al. [2008])	0.705942
Cox PH	0.666684

Table 3.1: Concordance achieved on synthetic datasets.

3.6.2 Real Data

We evaluate our approach compared to the state of the art on several open healthcare-related datasets. The four medical datasets considered are the Worcester Heart Attack Study (WHAS Hosmer, Lemeshow, and May [2000]), the Study to Understand Prognosis's Preferences Outcomes and Risks of Treatment (SUPPORT Knaus et al. [1995]), The Molecular

Dataset	$\mathbb{E}[\delta]$	d	n_{train}	n_{test}
Metabric	0.579307	9	1523	381
RGBSG	0.567652	7	1546	686
Support	0.680266	14	7098	1775
WHAS	0.421245	6	1310	328

TABLE 3.2: Descriptive statistics of the real datasets used in this work.

Taxonomy of Breast Cancer International Consortium (METABRIC Curtis et al. [2012]) as well as the Rotterdam & German Breast Cancer Study Group (RGBSG Foekens et al. [2000]; Schumacher et al. [1994]). The characteristics of the different datasets are summarized in table 3.2

The networks parameterizing σ_θ, π_θ, f_θ are chosen to be simple feed forward neural networks, but as π_θ is time independent, a more expressive network could be chosen such as a transformer network (Vaswani et al. [2017]) with a dense last layer if the covariates include unstructured text. Solving the neural differential equation involves repeated evaluation of the functional defining the dynamics; while solvers such as `Tsit5` (Tsitouras [2011]) and `BS5` (Bogacki and Shampine [1989]) are adaptive and only require a limited number of evaluations depending on the stiffness of the problem, we still keep the computational complexity of the time components manageable in order to keep training times low. As the accuracy of the solution is not of the utmost importance in our application,[105] we found it possible to use low accuracy solvers with high tolerances without any loss of predictive performance.

105: We are only interested in the generalization error, not the approximation error.

The parameters used for the survival normalizing flows are summarized in §3.6.2 (with L the number of layers and S their size), while the parameters and results from the other techniques are taken as-is from their respective papers. The activation function used is `SELU` (Klambauer et al. [2017]) for all datasets and we use the identity function as a last layer for the covariate networks π and the softmax function for the σ and θ. The performance

Dataset	S_π	S_σ	S_g	L_π	L_σ	L_g	K
Support	4	4	12	3	3	3	16
WHAS	12	8	12	4	4	3	32
RGBSG	4	4	12	3	3	3	16
Metabric	4	4	12	3	4	4	32

TABLE 3.3: Hyperparameters selected for the survival flows used in table 3.4.

of the different methods on the four datasets is summarized in table 3.4. We see that Normalizing Flows outperform the state-of-the-art on 2 of the 4 datasets and only underperform compared to random forests on the WHAS dataset. Such a result is not surprising: the covariates include highly engineered binary variables that were strongly suspected to be indicators of future heart problems by the instigators, it is therefore expected (provided that their hypothesis was correct) that a space partitioning algorithm

Method	Concordance			
	Support	WHAS	RGBSG	Metabric
This work	0.61678	0.86059	**0.68464**	**0.64879**
DeepSurv (Katzman et al. [2018])	**0.61831**	0.86262	0.66840	0.64337
RSF (Ishwaran, Kogalur, et al. [2008])	0.61302	**0.89362**	0.65119	0.62433
Cox PH	0.58287	0.81762	0.65775	0.63062

Table 3.4: Concordance of survival flows compared to competing techniques achieved on multiple real datasets.

would perform close to optimally.

3.7 Conclusion

We have shown how to build highly flexible generative models of the survival based on continuous normalizing flows and demonstrated that they are able to outperform state-of-the art techniques in exchange of a potentially sizeable increase in computational complexity. Nevertheless, the performance gain on the examples presented can justify in many cases the increased complexity since the majority of the computational power is used during the training phase and then amortized during inference, which is not the case for example in the case of k-NN or kernel estimators.[106] We will also show in chapter 5, using a toy example designed to resemble a real-world problem commonly faced in finance, that the ability to efficiently sample from the learned distribution can be incredibly valuable and more than offset the computational cost incurred during the learning phase. While the training wall time[107] of our model can probably be significantly improved by carefully optimizing the code up to the standard of the mature competing libraries, we still believe that more research is necessary in order to achieve the best possible performance. It is known in the regression setting that augmenting the ODE (Dupont, Doucet, and Teh [2019]) leads to significantly improved predictive as well as computational performance, but those results cannot be directly applied to the normalizing flow setting. Similarly, methods that try to control the stiffness of the ODE such as Finlay et al. (2020) have shown great promises.

Additionally, a more direct and potentially simpler way of reducing the computational complexity is simply to reduce the complexity of the conditioning, and therefore the size of the neural networks involved in the parametrization of the flow. *Variable selection*, is a simple method for reducing the number of dimensions of the covariates that can benefit most models and is particularly appealing when the number of covariates is overwhelming as can be the case with financial or genetic datasets. However, variable selection is usually treated as a separate problem from the downstream task of prediction and therefore may select dimensions that are important in general[108] but useless for the task of prediction and

106: Both k-NN and kernel type estimators needs to be "retrained" for every new observation. Not only that but inference cost scales with the size of the training set, a significant downside.

107: Elapsed real time.

108: For some definition of *important*, such as explainability of the variance in the case of PCA.

in our specific case, censored prediction. In the next chapter we will therefore show, amongst other things, how to select variables that are useful for prediction in the survival setting by means of estimating the gradient of the regression function and only selection the dimensions with a non-zero gradient under the assumption that useful variables are variables that impact the output i.e. the regression function.

Prediction in High Dimension | 4

4.1 Introduction

Advances in computing power as well as storage capacity have led to an exponential increase in the quantity of data generated and captured each year. It is not surprising that in this fertile environment, machine learning and *big data*[110] have developed rapidly as important fields pushing organizations to accumulate as much data as possible, be it every text reports on a company in the banking environment, all the electronic recordings of the various monitoring instruments in the medical setting or more generally all the data one can hope to acquire.

110: We define *big data* as techniques to process data too voluminous to fit on a single machine. Contrary to popular corporate belief this is *far more* data than any Excel sheet can hold. A single modern server can hold several terabytes of data in volatile memory and hundred of terabytes of data on disk.

4.1.1 The Curse of Dimensionality

While this volume of data has largely been useful and responsible for the surge of new applications of machine learning; practitioners have quickly found that while many *observations* is always a good thing, observations of many characteristics are not necessarily. It is not rare that adding more data about individual observations not only do not improve the quality of the models as expected, but worse can result in worse predictions. This curious, at first glance, phenomenon is not new to theorists and researchers of many fields. First coined as the *curse of dimensionality* by Richard E. Bellman in the setting of dynamic programming (Bellman [1954]), refer to the seemingly inescapable negative influence of too many covariates when at the same time the necessity to add more data seems inevitable. In chapter 2, the main results given expose their exposure to the curse of dimensionality through the dependence of the bounds Proposition 2.8 and Theorem 2.9 on d the dimension of $\mathbb{X}$. While mostly a simplification, it is possible to understand why the highly dimensional setting is hard by visualizing the size of the neighbourhoods for d large.[111] As d grows, the volume of unit balls grows exponentially and consequently the proportion of the space needed to encompass a certain percentage of the observations grows quickly. Intuitively, as the dimension grows the density of neighbourhoods decrease and as those become depleted it becomes necessary to use larger neighbourhoods to compensate. Similarly, while not explicitly written

Figure 4.1: The IBM 305 RAMAC introduced in 1956 and weighing 1 Ton could hold 5 MB for the monthly price of $30.000.

111: This is in this case largely the scheme followed in the proofs of chapter 2 where we build neighbourhoods defined by the metric induced by the kernel, k-NN or tree.

down, the performance of the normalizing flow methods of chapter 3 depend greatly on the number of covariates, not only from an estimation point of view because of the risk of overfitting but more problematic, in this case, for purely computational reasons. Given those observations, it is clear that the ability to reduce the effective dimension of the input space intelligently is a worthwhile endeavour.

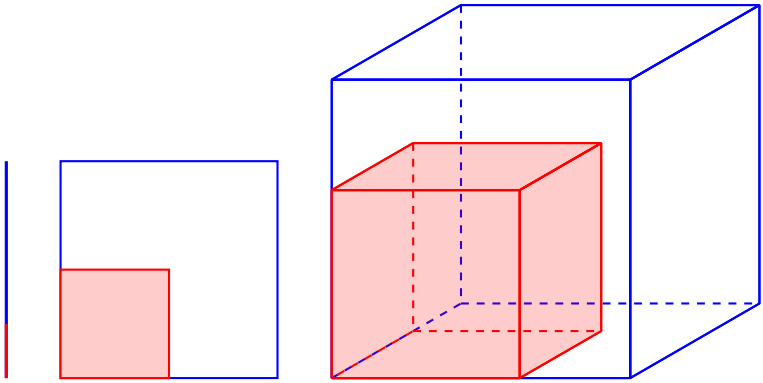


FIGURE 4.2: A neighbourhood representing ¼ of the space for dimensions 1, 2 and 3.

4.1.2 Reduction of Dimension

The goal of dimensionality reduction is the find a *new space* with fewer dimensions than the original space but the same amount of *information*.[112] Given this fairly vague definition it is not surprising that it is possible to derive many different techniques to achieve this goal, depending on the definitions of *new space* and *information* one decides to adopt. Given that our goal is not specifically to study the vast field of dimension reduction but only one specific example adapted to survival analysis, we take the liberty of very roughly dividing those techniques in two large families: *supervised* and *unsupervised* dimension reduction.

112: Otherwise dimensionality reduction is easy: just drop random covariates.

Unsupervised[113] dimension reduction will refer here to methods that try to find new smaller representations of $\mathbb{X}$ independently of any other potential downstream tasks. The most famous representant of this family being the time honored PCA of Pearson (1901) which can be seen either as a lossy reconstruction method where one try to find the projection of dimension k that minimizes the square reconstruction error, or from a probailistic point of view where, after some hypothesis on the underlying graphical model, one try to minimize the information loss, that is maximize the mutual information between the full data and the reduced dimension data. The first approach can then be extended naturally to handle sparsity (Zou, Hastie, and Tibshirani [2006]) or even non-linearities (Lu, Plataniotis, and Venetsanopoulos [2011]; Hastie and Stuetzle [1989]; Schölkopf, Smola, and

113: Some people prefer the term *self-supervised*, to make clear that the covariates also play the role of the target variable.

K.-R. Müller [1998]; Gorban et al. [2008]) while the second view has been extended to give the proper probabilistic principal component analysis (PPCA) (Tipping and Bishop [1999]). More generally most methods try to preserve some intrinsic notion of the geometry of the data: multidimensional scaling methods (M. A. A. Cox and T. F. Cox [2008]) strive to preserve dissimilarities between observations, locally linear embedding (LLE) (Roweis and Saul [2000]) attempts to preserve the local geometry while stochastic neighbor embedding (SNE) and uniform manifold approximation and projection for dimension reduction (UMAP) (McInnes, Healy, and Melville [2020]; Hinton and Roweis [2003]) trie to perserve the neighbourhoods of the global space in the lower dimensional embedding. More recently, approaches based on extracting the rich representations from deep neural networks have enjoyed great successes by profitting from the rapid advances of the field. While the end result of those approaches is in many way similar: obtaining lower dimensional representations[114] that retain most of the information about the original dataset, the fact that those representations are highly processed compared the fairly mild usual dimensionality reduction techniques[115] and are specific to each dataset and learned end-to-end, have resulted in those techniques forming their own subfield under the taxonomy of *representation learning* (Y. Bengio, Courville, and Vincent [2014]). VAE (Kingma and Welling [2014, 2019]), the main well principled representant of this field adopts the ideas of the previously given examples, that is minimizing some reconstruction loss[116] in order to learn a domain specific mapping resulting in a good lower dimensional embedding of the input space. While these increasingly complex approaches to building representations have proven incredibly successful,[117] simpler methods such as random projections (Johnson, Lindenstrauss, and Schechtman [1986]) and more recently random fourier features (Rahimi and Recht [2008]) are still competitive and worthwhile tools in the dimension reduction toolbet of every practitioner.

Supervised dimension reduction on the other hand makes use of auxiliary data in order to guide the reduction, usually by taking into account the downstream task i.e. the target variable Y in the case of regression. The general framework is mostly similar to the unsupervised case: trying to find a low dimension representation that retains most of the useful information. However, while in the unsupervised setting the notion of *information* has to be defined a-prori and can be seen as *information about itself*, in the supervised setting it is possible to define it more straighforwardly and ad-hoc as in this case information refers directly to information useful to the task at hand. As such, while PCA tried to find some linear transformation that maximized variance, its supervised complement, LDA[118] (Fisher [1936]) tries to find a linear mapping that maximally separate classes,[119] therefore making use of the target as guide. Similarly to PCA, LDA has been extended

114: Or *embeddings* in the deep-learning litterature.

115: Which often are full subspaces or projections of the input space.

116: Seen here from a probabilistic loss as a lower bound of the evidence, or evidence lower bound (ELBO). VAE are intimately linked to the notion of normalizing flow (NF) presented in chapter 3.

117: The recent successes in NLP can largely be credited to the success of representation learning, Mikolov et al. (2013); Pennington, Socher, and Manning (2014), see.

118: Not to be confused with latent dirichlet allocation.

119: When the target is a classification variable.

to non-linearities (Baudat and Anouar [2000]). Previously we have given the example of multiple techniques which strived to preserve neighbourhoods in the new lower dimensional space, by analogy the same techniques can be applied in order to preserve *class neighbourhood* (Salakhutdinov and Hinton [2007]), in order to exploit the class information. Analoguously to representation learning, *feature learning*, has proven popular in the field of deep-learning where intermediary outputs[120] are taken as meaningful task specific representations of the covariates. Another possible approach, and the one we will take in this chapter, is to consider the gradient of the regression function or loss of interest with respect to the covariates as a measure of feature importance. If the task of interest can be expressed as a regression $\mathbb{E}[(Y - f(X))^2]$ then $\nabla f(x)$ represents the contribution of each specific dimension to the result. It seems therefore natural to consider the directions with nearly zero directional derivatives to be irrelevant and discard them. Note that in this case the definition of variable or direction of importance is then local at a neighbourhood of x. While it is possible to obtain a global view by averaging over all the possible x for example, it can be more beneficial to use the locality in our favor. We will give examples of this specific fact later in §4.5. Because the function f itself is unknown, the gradient too usually has no hope of being knowable and therefore has to be estimated in order to compute some variable importance metric based on the gradient. As the empirical gradients retrieved have no reasons to have any zero dimensions, Y. Kim and J. Kim (2004); Ye and Xie (2012) propose sparse formulations based on the LASSO. Sheth and Fusi (2019) on the other hand propose a relaxation method to deal with large gradient learning problems. Those methods however are mainly concerned with the solvability of the gradient learning problem from a computational point of view and not with the quality of the estimate as well as the generalization error of the final task. L. Yang et al. (2017) do give asymptotic results when the underlying model is assumed to be partially linear but as in chapter 2 we do not wish to make any assumptions on the form of the true f and would like to obtain non asymptotic results.

As the empirical gradient is a useful proxy for variable importance as well as a useful object in its own right, we will study in this chapter the problem of *empirical gradient estimation*, and try to give results similar as those of chapter 2, that is non-parametric and nonasymptotic learning bounds for the underlying learning problem. We will show how the resulting gradient estimate can be used for variable selection and adapted to the survival setting, as well as for other tasks involving the gradient.

120: Usually the penultimate layer of a deep network.

4.2 Empirical Gradient Estimation

About this Section

The rest of this chapter is in large part reproduced from the paper "Nearest neighbour based estimates of gradients: Sharp nonasymptotic bounds and applications" with Stéphan Clémençon and François Portier published in the proceedings of AIStats'21.

Guillaume Ausset, Stéphan Clémençon, and François Portier (2021b). "Nearest Neighbour Based Estimates of Gradients: Sharp Nonasymptotic Bounds and Applications". In: *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics*. Ed. by Arindam Banerjee and Kenji Fukumizu. Vol. 130. Proceedings of Machine Learning Research. PMLR, pp. 532–540

In this section, we place ourselves in the same regression setup as introduced and studied in chapter 2 but without any censoring. Here and throughout the rest of the chapter, (X, Y) is a pair of random variables defined on the same probability space $(\omega, \mathcal{F}, \mathbb{P})$ with unknown probability distribution $\mathbb{P}$. The random variable Y is real valued and supposed square integrable, whereas the supposedly continuous random vector X takes its values in $\mathbb{R}^d$, with $d \geq 1$, and models some information *a priori* useful to predict Y. Based on a sample $\mathcal{D}_n$ of $n \geq 1$ independent copies of the pair (X, Y), i.e.

$$\mathcal{D}_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\},$$

the goal pursued is to build a Borelian mapping $f : \mathbb{R}^d \to \mathbb{R}$ that produces, on average, a good prediction $f(X)$ of Y. Measuring classically its accuracy by the squared error, the learning task then boils down to finding a predictive function f that is solution of the risk minimization problem $\min_f \mathcal{R}(f)$, where

$$\mathcal{R}(f) = \mathbb{E}\left[(Y - f(X))^2\right]. \tag{4.1}$$

Of course, as discussed at length in chapter 2, the true minimum which is attained by the regression function $r(X) = \mathbb{E}[Y \mid X]$ is unknown as is the conditional distribution of Y given X and the population risk eq. (4.1). Following the ERM strategy of chapter 2, which consists in solving the optimization problem above, but with the unknown distribution $\mathbb{P}$ replaced by an empirical estimate $\mathbb{P}_n$ based on the training data $\mathcal{D}_n$, such as for example the raw empirical distribution

$$\mathbb{P}_n = \frac{1}{n} \sum_{i \leq n} \delta_{X_i},$$

where δ_x is the the Dirac distribution at some point x. The resulting ERM problem is then minimized on some restricted class $\mathcal{F}$ supposed to be rich enough to include a reasonable approximant of f but not too complex (e.g. of finite VC dimension) in order to control the fluctuations of the deviations between the empirical and true distributions uniformly over it. As shown in §2.1, under the assumption that the random variables Y and $f(X)$ for $f \in \mathcal{F}$ have sub-Gaussian tails, the analysis of the performance of empirical risk minimizers (i.e. predictive functions obtained by least-squares regression) can be controlled and provides nonasymptotic and nonparametric bounds on the error. For a better view of the ERM perspective and what type of generalization bounds we encourage the reader to go back to §2.1 for the uncensored case and §2.2 for the censored setting.

In this section, we are interested in estimating accurately the supposedly well-defined gradient $\nabla r(x)$ by means of the popular k-NN approach (see e.g. Devroye, Györfi, and Lugosi [1996], chapter 11; or Biau and Devroye [2015]). The *gradient learning* issue has received increasing attention in the context of local learning problems such as classification or regression these last few years (see e.g. Mukherjee and Wu [2006]; Mukherjee and D.-X. Zhou [2006]). Because it provides a valuable information about the local structure of a dataset in a high-dimensional space, an accurate estimator of the gradient of a predictive function can be used for various purposes such as dimensionality reduction or variable selection, the partial derivative with respect to a given variable being a natural indicator of its importance regarding prediction. In Hristache, Juditsky, and Spokoiny (1998), the authors study the *single index* regression model, that is assume that the useful explanatory data is contained in a 1-dimensional subspace obtained by projection of the space $\mathbb{X}$, that is

$$Y_i = g(TX_i) + \varepsilon_i,$$

where $g : \mathbb{R} \mapsto \mathbb{R}$ is called the link function and the space spanned by $v \mapsto Tv$ is called the EDR space.[121] They extend this setting in Hristache, Juditsky, Polzehl, et al. (2001) to the multi-index setting, that is the same setting but with $g : \mathbb{R}^k \mapsto \mathbb{R}$ and therefore T an orthonormal projection of rank k. In this specific setting, one can straighforwardly prove that the gradient ∇f of the true regression function f belongs to the index space at every point X_i. The authors therefore propose to estimate the matrix

$$\mathbb{M} = \frac{1}{n} \sum_{i=1}^{n} \nabla f(X_i) \nabla^\intercal f(X_i), \tag{4.2}$$

in order to find the principal components of $\mathbb{M}$, that is

$$\mathbb{M} = O_d^\intercal \Lambda O_d,$$

121: *Single* index because here of dimension 1. Extensions to the *multi index* setting i.e. for dimensions of the subspace more than 1 also exist.

where O_d is an orthonormal matrix and Λ is a diagonal matrix of decreasing eigenvalues. While the true $\mathbb{M}$ is unknown, if one is able to estimate the gradient of the regression function then taking the k largest components of the decomposition of

$$\hat{\mathbb{M}}_n = \frac{1}{n}\sum_{i=1}^{n} \hat{\nabla}_n f(X_i)\hat{\nabla}_n^\intercal f(X_i),$$

yields an estimate of the k indexes subspace and therefore provided that the model is correct a good dimensionality reduction method. Several improvements to this procedure have been proposed in order to improve the quality of the estimator. In Dalalyan, Juditsky, and Spokoiny (2008), the authors propose to replace eq. (4.2) by

$$\mathbb{M}_L = \sum_{i=1}^{L} \beta_i \beta_i^\intercal, \tag{4.3}$$

where the β_i are defined by

$$\beta_k = \frac{1}{n}\sum_{i=1}^{n} \nabla f(X_i)\psi_k(X_i),$$

with the ψ_k forming an orthogonal basis such that

$$\frac{1}{n}\sum_{i=1}^{n} \psi_l(X_i)\psi_m(X_i) = \delta_{l,m}.$$

In this case, the eigenvectors of $\mathbb{M}$ are eigen vectors of $\mathbb{M}_L$ and $\mathbb{M}_n = \mathbb{M}$. The authors show that $\mathbb{M}_L$ is easier to estimate and that the estimated projection matrix formed from the obtained eigenvectors can be written as

$$\begin{aligned} \hat{T}_n &\in \operatorname*{argmin}_{T\in\mathcal{A}_m} \max_l \ \hat{\beta}_l^\intercal \left(\mathbb{I} - T\right)\hat{\beta}_l, \\ \text{s.t.} \quad \mathcal{A}_m &= \left\{T : T = T^\intercal,\ 0 \preceq T \preceq \mathbb{I},\ \operatorname{tr} T \leq m\right\}, \end{aligned}$$

where m is the guessed EDR dimension. The previous references are all concerned with outer products of gradients so as to recover some dimension-reduction subspace. Estimators of the gradients have also been proposed for zeroth-order optimization (Nesterov and Spokoiny [2017]; Yining Wang et al. [2018]; Berahas et al. [2020], see e.g.) and can benefit from good convergence properties.Another possible approach is to remark that the quantity of interest in eq. (4.3) is not actually the gradient but the expected gradient outerproduct (EGOP) $\mathbb{E}[\nabla f(X)\nabla^\intercal f(X)]$. While estimating the gradient first in order to build a plugin estimator of this outerproduct

is, of course, a perfectly viable direction; it is, if at all possible, a better idea to estimate directly the quantity of interest without relying on plugin estimators as it is easier to control the risk of a simple estimator than the product of estimators. This is the approach taken by Xia et al. (2002); Xia (2007); Trivedi et al. (2014), where kernel estimators of the gradient outerproduct are proposed and shown to be asymptotically consistent. Ye and Xie (2012) extend these kernel estimators to exploit the potential sparsity of the gradient by use of a LASSO formulation in order to give non-asymptotic results on the quality of the estimator of the gradient. Whereas the use of standard nonparametric methods for gradient estimation is documented in the literature (see Fan and Gijbels [1996]; Delecroix and Rosa [1996]; De Brabanter et al. [2013], for the use of local polynomial with kernel smoothing techniques; Gasser and H.-G. Müller [1984], for the so-called Gasser-Muller alternative; S. Zhou and Wolfe [2000], for the use of regression spline; and Mukherjee and D.-X. Zhou [2006], for the estimation on a reproducing kernel Hilbert space with kernel smoothing), it is the purpose of the rest of this chapter to investigate the performance of an alternative local averaging method, the popular k-NN method. As k-NN provides piecewise constant estimates, it is easier to conceptualize for the practitioner and, more importantly, the neighbourhoods determined by the parameter k are data-driven and often more consistent than those defined by the bandwidth in the kernel setting, especially in high dimensions. Whereas one has to adapt the bandwidth of the kernels in the kernel averaging setting in order to ensure a certain number of points are always contained in the implicitly defined ball, the risk of pathological cases where some points are isolated from the rest grows with the dimension. This problem is entirely alleviated by the choice of k-NN, by definition, at the cost of neighbourhoods of varying sizes which therefore makes the theoretical analysis less straightforward.

Here we investigate the behaviour of the estimator of the supposedly sparse gradient of the regression function at a given point $x \in \mathbb{R}^d$, obtained by solving a regularized local linear version of the k-NN problem with a Lasso penalty. Precisely, nonasymptotic bounds for the related estimation error are established. Whereas k-NN estimators of the regression function have been extensively analysed (see e.g. Biau, Cérou, and Guyader [2010]; Kpotufe [2011]; Jiang [2019], and the references therein), the result stated in this section is the first combining both uniform asymptoticity as well a nearest neighbour approach, to the best of our knowledge.

While the dimension reduction aspect has already been exposed in detail earlier in this section, the relevance of the approach promoted is illustrated in §4.5 by several applications, even beyond the primary goal of dimension reduction. Reduction dimension is first studied in §4.5.1 for the standard regression setting where variable selection algorithm

that exploits the local nature of the gradient estimator proposed is first exploited to refine the popular random forest algorithm (see Breiman [2001]): by exploiting the node estimate of the gradient we are able to direct better the choice of cuts. Very simple to implement and accurate, as supported by the various numerical experiments carried out, it offers an attractive and flexible alternative to existing traditional methods such as PCA or the more closely related method of Dalalyan, Juditsky, and Spokoiny (2008), allowing for a local reduction of the dimension rather than implementing a global preprocessing of the data. We then turn our attention to the censored setting in §4.5.2 where the survival gradient is estimated in order to retrieve the genes relevant to the prediction of cancer. We next show in §4.5.3 how a rough statistical estimate of the gradient of any smooth objective function based on the estimation principle previously analysed in the context of regression can be exploited in a basic gradient descent algorithm. We exploit the local structure of the algorithm to be able to reuse past computations in order to calculate our estimator and jump to a better local minimum at each gradient step as well. Finally in §4.5.4, we give an example of the usefulness of a sparse gradient estimate when the gradient is believed to be truly sparse: we use our estimator to retrieve the direction of interest for a specific attribute inside a disentangled representation and show how this can be used as an *ad-hoc* measure of disentanglement.

4.3 Sparse Local Linear Regression

Before stating the main theoretical results in §4.4, we start by quickly exposing the theoretical setting and assumptions required for the results, as well as replace our approach among the previous results in the litterature. We postpone the technical proofs to §4.7 at the end of the chapter in order not to hinder readability.

As we rely heavily on the definition of neighbourhood and therefore their associated metrics, we start by giving a reminder of the notations we will use for the various norms necessary in the proofs and results. For any vector $x = (x_1, \dots, x_d)$ in $\mathbb{R}^d$ we write the ℓ_∞-norm, the ℓ_1-norm and the ℓ_2-norm as

$$\begin{aligned} \|x\|_\infty &\stackrel{\text{def}}{=} \max\Big(|x_1|, \dots |x_d|\Big), \\ \|x\|_1 &\stackrel{\text{def}}{=} |x_1| + \dots + |x_d|, \\ \|x\|_2 &\stackrel{\text{def}}{=} \sqrt{x_1^2 + \dots + x_d^2}. \end{aligned}$$

We also define by $\mathcal{B}(x, \tau)$ the closed ball of centre $x \in \mathbb{R}^d$ and radius $\tau > 0$,

that is

$$\mathcal{B}(x,\tau) \stackrel{\text{def}}{=} \left\{z \in \mathbb{R}^d : \|x - z\|_\infty \leq \tau\right\}.$$

4.3.1 k-NN estimation methods in regression

Let $x \in \mathbb{R}^d$ be fixed and $k \in \{1,\ldots,n\}$. Define

$$\hat{\tau}_k(x) \stackrel{\text{def}}{=} \inf\left\{\tau \geq 0 : \sum_{i=1}^{n} \mathbb{1}_{\mathcal{B}(x,\tau)}(X_i) \geq k\right\},$$

which quantity is referred to as the k-NN radius. Indeed, observe that, equipped with this notation, $\mathcal{B}(x,\hat{\tau}_k(x))$ is the smallest ball with centre x containing k points of the sample $\mathcal{D}_n$ and the mapping $\alpha \in (0,1] \mapsto \hat{\tau}_{\alpha n}(x)$ is the empirical quantile function related to the sample $(\|x - X_1\|_\infty)_i$. The rationale behind k-NN estimation in the regression context is simplistic, the method consisting in approximating the regression function $r(x) = \mathbb{E}[Y \mid X = x]$ by $\mathbb{E}[Y \mid X \in \mathcal{B}(x,\tau)]$, the mapping r being assumed to be smooth at x, and computing next the empirical version of the approximant (i.e. replacing the unknown distribution $\mathbb{P}$ by the raw empirical distribution). This yields the estimator

$$\hat{r}_k(x) = \frac{1}{k}\sum_{i \in \hat{i}_k(x)} Y_i, \tag{4.4}$$

usually referred to as the standard k-nearest neighbour predictor at x and where we define $i(x)$ as the neighbourhood

$$\hat{i}_k(x) = \left\{j : X_j \in \mathcal{B}(x,\hat{\tau}_k(x))\right\}.$$

Of course, the mapping $x \in \mathbb{R}^d \mapsto \hat{r}_k(x)$ is locally/piecewise constant, just like $x \in \mathbb{R}^d \mapsto \hat{\tau}_k(x)$. The local average $\hat{r}_k(x)$ can also be naturally expressed as

$$\hat{r}_k(x) = \operatorname*{argmin}_{r \in \mathbb{R}} \sum_{i \in \hat{i}_k(x)} (Y_i - r)^2. \tag{4.5}$$

For this reason, the estimator eq. (4.4) is sometimes referred to as the *local constant* estimator in the statistical literature. Following in the footsteps of the approach proposed in Fan (1992), the estimation of the regression function at x can be refined by approximating the supposedly smooth $r(z)$ around x in a linear fashion, rather than by a local constant m, since we have by virtue of a first-order Taylor expansion:

$$r(z) = r(x) + \nabla r(x)^\top (z - x) + o(\|z - x\|).$$

For any point X_i close to x, one may write locally

$$r(X_i) \approx r + \beta^\top (X_i - x)$$

and the *local linear* estimator of $r(x)$ and the related estimator of the gradient $\beta(x) = \nabla r(x)$ are then defined as

$$\underset{(r,\beta)\in\mathbb{R}^{d+1}}{\operatorname{argmin}} \sum_{i\in\hat{i}_k(x)} \left(Y_i - r - \beta^\intercal (X_i - x)\right)^2. \tag{4.6}$$

Because of its reduced bias, the local linear estimator (the first argument of the solution of the optimization problem above) can improve upon the local constant estimator eq. (4.4) in moderate dimensions. However, when the dimension d increases, its variance becomes large and the design matrix of the regression problem is likely to have small eigenvalues, causing numerical difficulties. For this reason, we introduce here a lasso-type regularized version of eq. (4.6), namely

$$\underset{(r,\beta)\in\mathbb{R}^{d+1}}{\operatorname{argmin}} \sum_{i\in\hat{i}_k(x)} \left(Y_i - r - \beta^\intercal (X_i - x)\right)^2 + \lambda\|\beta\|_1, \tag{4.7}$$

with solution[122] $(\tilde{r}_k(x), \tilde{\beta}_k(x))$ and where $\lambda > 0$ is a tuning parameter governing the amount of ℓ_1-complexity penalization. For the moment, we let it be a free parameter and will propose a specific choice in the next section. Focus is here on the gradient estimator $\tilde{\beta}_k(x)$, i.e. the second argument in eq. (4.7). In the subsequent analysis, nonasymptotic bounds are established for specific choices of λ and k. The following technical assumptions are required.

122: Note that the problem depends on x and that the solution is therefore a function of x.

4.3.2 Technical hypotheses

The hypothesis formulated below permits us to relate the volumes of the balls $\mathcal{B}(x,\tau)$ to their probability masses, for τ small enough.

Assumption 4.1. There exists $\tau_0 > 0$ such that restriction of X's distribution on $\mathcal{B}(x,\tau_0)$ has a bounded density g, bounded away from zero, with respect to Lebesgue measure:

$$b_f = \inf_{y\in B(x,\tau_0)} g(y) > 0,$$
$$U_f = \sup_{y\in B(x,\tau_0)} g(y) < +\infty.$$

Suppose in addition that $U_f/b_f \leq 2$.

The constant 2 involved in the condition above for notational simplicity can be naturally replaced by any constant $1+\gamma$, with $\gamma > 0$. The next assumption, useful to control the variance term, is classical in regression, it stipulates that we have $Y = r(x) + \varepsilon$, with a sub-Gaussian residual ε independent from X.

Assumption 4.2 (Sub-Gaussianity)**.** The zero-mean and square integrable random variable $\varepsilon = Y - r(x)$ is independent from X and is sub-Gaussian with parameter $\sigma^2 > 0$, i.e. $\forall \lambda \in \mathbb{R}$,

$$\mathbb{E}\left[\exp(\lambda\varepsilon)\right] \leq \exp\left(-\frac{\sigma^2\lambda^2}{2}\right),$$

In order to control the bias error when estimating the gradient $\beta(z) = \nabla r(z)$ of the regression function at x, smoothness conditions are naturally required.

Assumption 4.3 (Residual lipschitzianity)**.** The function $r(z)$ is differentiable on $\mathcal{B}(x, \tau_0)$ with gradient $\beta(z) = \nabla r(z)$ and there exists $L_2 > 0$ such that for all $z \in \mathcal{B}(x, \tau_0)$,

$$\left|r(z) - r(x) - \beta(x)(z - x)\right| \leq L_2 \left\|z - x\right\|_\infty^2 .$$

Finally, a Lipschitz regularity condition is required for the density g.

Assumption 4.4 (g lipschitzianity)**.** The function g is L-Lipschitz at x on $\mathcal{B}(x, \tau_0)$, i.e. there exists $L > 0$ such that for all $z \in B(x, \tau_0)$,

$$\left|g(z) - g(x)\right| \leq L \left\|z - x\right\|_\infty .$$

We point out that, as the goal of this chapter is to give the main ideas underlying the use of the k-NN methodology for gradient estimation rather than carrying out a fully general analysis, the ℓ_∞-norm is considered here, making the study of ℓ_1 regularization easier. The results of this chapter can be extended to other norms at the price of additional work.

4.4 A k-NN based estimator of the gradient

The main theoretical result of the present paper is now stated and further discussed. Under the hypotheses listed in the previous section and for specific choices of k and λ, it provides a nonasymptotic bound for the estimator $\tilde{\beta}_k(x)$ of the gradient $\beta(x) = \nabla r(x)$ at x given by eq. (4.7). Whereas nonasymptotic bounds for k-NN estimators of the regression function have been established under various smoothness assumptions (see e.g. Jiang [2019]; or Kpotufe [2011]), no nonasymptotic study of k-NN based estimator of the gradient of the regression function is documented in the literature. To the best of our knowledge, the result proved in this article is the first of this nature. Two key quantities are involved in the upper confidence bound given in Theorem 4.1, the (deterministic) radius

$$\overline{\tau}_k = \left(\frac{2k}{nb_f 2^d}\right)^{1/d},$$

that upper bounds the k-NN radius on an event holding true with high probability, with corresponding neighbourhood

$$\bar{\imath}_k(x) = \left\{ j : X_j \in \mathcal{B}(x, \overline{\tau}_k(x)) \right\},$$

as well as the cardinality of the so-called local active set

$$\mathcal{S}_x = \left\{ 1 \leq k \leq d : \beta_k(x) \neq 0 \right\},$$

which, for clarity reasons, is supposed to be non-empty.

Theorem 4.1. *Suppose that Assumptions 4.1 to 4.4 are fulfilled. Let $n \geq 1$ and $k \geq 1$ such that $\overline{\tau}_k \leq \tau_0$.*

$$\lambda = \overline{\tau}_k \left(\sqrt{2\sigma^2 \frac{\log(16d/\delta)}{k}} + L_2 \overline{\tau}_k^2 \right).$$

Then, we have with probability larger than $1 - \delta$,

$$\left\| \tilde{\beta}_k(x) - \beta(x) \right\|_2 \leq 24^2 \sqrt{|\mathcal{S}_x|} \left(\overline{\tau}_k^{-1} \sqrt{\frac{2\sigma^2 \log(16d/\delta)}{k}} + L_2 \overline{\tau}_k \right), \quad (4.8)$$

as soon as

$$C_1 \, |\mathcal{S}_x| \log\left(\frac{dn}{\delta} \right) \leq k \leq C_2 n,$$
$$\overline{\tau}_k^2 \leq \frac{b_f^2}{C_3 \, |\mathcal{S}_x| \, L^2} \wedge \tau_0^2,$$

where C_1, C_2 and C_3 are universal constants.

The analysis of the accuracy of the nearest neighbour estimate $\hat{r}_k(x)$ classically involves the following decomposition of the estimation error

$$\hat{r}_k(x) - r(x) = (\hat{r}_k(x) - r_k(x)) + (r_k(x) - r(x)), \quad (4.9)$$

where

$$r_k(x) = \frac{1}{k} \sum_{i \in \hat{\imath}_k(x)} r(X_i).$$

The approach developed in Jiang (2019) essentially consists in combining this decomposition with the fact that $\hat{\tau}_k(x) \leq \overline{\tau}_k$ with high probability. By its own nature, our local linear Lasso regularized estimate of the gradient $\tilde{\beta}_k$ cannot be treated in the same way. First, in order to take advantage of the Lasso regularization in sparse situations (i.e. when the gradient at x depends on a small number of covariates solely), we rely on a basic

inequality from Hastie, Tibshirani, and Wainwright (2015), Lemma 11.1 which is useful when analysing standard Lasso estimates. Second, we need to control the size of the neighbourhoods $\hat{\tau}_k(x)$ on an event of high probability. In this respect, we slightly deviate from the approach of Jiang (2019): we do not rely on concentration results over VC classes but only on the Chernoff concentration bound. This way, we can relax significantly the lower bound conditions for k as the dimension d increases, see Theorem 4.2 below, which compares favourably with Corollary 1 in Jiang (2019) for instance.

Balancing between the bias and the variance term of the upper bound provided in eq. (4.8) we obtain that the optimal value for k is $k \sim n^{4/(4+d)}$. In this case, the bound stated above yields the rate $n^{-1/(4+d)}$. As a consequence, our bound matches the minimax rate (up to log terms) given in Stone (1982) for the problem of the estimation of the derivative (in a L_2 sense).

4.4.1 Pointwise k-NN estimation of $r(x)$

Though it concerns the local estimation error, the bound in Theorem 4.2 below can be viewed as a refinement of the nonasymptotic results recently established in Jiang (2019) (see also Kpotufe [2011]), which provide uniform bounds in x. It requires a local smoothness condition for the regression function. From now on, $\|\cdot\|$ denotes any norm on $\mathbb{R}^d$.

Assumption 4.5. The regression function $r(z)$ is L_1-Lipschitz at x, i.e. there exists $L_1 > 0$ such that for all $z \in \mathcal{B}(x, \tau_0) = \{x' \in \mathbb{R}^d : \|x' - x\| \leq \tau_0\}$,

$$|r(x) - r(z)| \leq L_1 \|x - z\| .$$

Theorem 4.2. *Suppose that Assumptions 4.1, 4.2 and 4.5 are fulfilled and that $2k \leq n\tau_0 b_f V_d$. Then for any $\delta \in (0,1)$ such that $k \geq 4\log(2n/\delta)$, we have with probability $1-\delta$:*

$$|\hat{r}_k(x) - r(x)| \leq \sqrt{\frac{2\sigma^2 \log(4/\delta)}{k}} + L_1 \left(\frac{2k}{n b_f V_d}\right)^{1/d}, \tag{4.10}$$

where $V_d = \int \mathbb{1}_{\mathcal{B}(0,1)}(x)\, \mathrm{d}x$ denotes the volume of the unit ball.

We obtain a weaker condition on the value of k than that obtained in Jiang (2019) (see Corollary 1 therein), due to our different treatment of the approximation term (the second term in decomposition eq. (4.9)) is different (see the argument detailed in §4.7). With $k \sim n^{2/(2+d)}$, the bound stated above yields the minimax rate $n^{-1/(d+2)}$.

4.5 Numerical Experiments

In order to motivate the need for a robust estimator of the gradient, we introduce three different examples of use of our estimator compared to existing approaches. All the code necessary for the reproduction of the experiments as well as figures can be found at `git.sr.ht/~aussetg/locallinear`.

As our estimator is sensitive to the choice of hyperparameters k and λ we use a local leave-one-out procedure described in algorithm 2 for hyperparameter selection. As only the regression variable Y is observed, the regression error is used as a proxy loss in the cross-validation. The high cost of k-NN is amortized by using k-d trees, bringing the total average complexity of the nearest neighbour search down to $O(n \log n)$. In cases where the aforementioned cost is too high (n in the order of millions) it is possible to instead make use of approximate nearest neighbour schemes such as HNSW (Malkov and Yashunin [2020]). Approximate Nearest Neighbours algorithms have recently enjoyed a regain of interest and provide high accuracy at a very low computational cost (Aumüller, Bernhardsson, and Faithfull [2018]).

Algorithm 2 Local Leave-One-Out

Require: x: sample point, (X, Y): training set, (K, λ): grid

$X_{\text{LoO}} \leftarrow$ `Neighbourhood of` x `in` X `of size` n
for $k \in K, \lambda \in \lambda$ **do**
  for $X_i \in X_{\text{LoO}}$ **do**
    $r_i, \beta_i \leftarrow$ `estimated gradient at` X_i `w.r.t` X, Y `using` eq. (4.7)
  end for
  $\texttt{error}_{k,\lambda} \leftarrow \frac{1}{n} \sum_{i=1}^{n} (r_i - Y_i)^2$
end for
$k^\star, \lambda^\star \leftarrow \operatorname{argmin}_{k,\lambda} \texttt{error}_{k,\lambda}$
return $k^\star, \lambda^\star$

4.5.1 Variable Selection

While a large number of observations is desirable the same is not necessarily the case for the individual features; a large number of features can be detrimental to the computational performance of most learning methods but also harmful to the predictive performance. In order to mitigate the detrimental impact of the high dimensionality, or *curse of dimensionality*, one can try to reduce the effective dimension of the problem. A large body of work exists on dimensionality reduction as a preprocessing step that considers the intrinsic dimensionality of X by considering for example

that X lies on a lower-dimensional manifold. Those approaches only consider X in isolation and do not take into account Y which is the variable of interest. It is possible to use the information in Y to direct the dimension reduction of X, either by treating Y as side information, as is done in Bach and Jordan (2005), or by considering the existence of an explicit *index space* such that $Y_i = g(v_1^\intercal X_i, \ldots, v_m^\intercal X_i) + \varepsilon_i$ as is done in Dalalyan, Juditsky, and Spokoiny (2008). In the latter case, it is possible to observe that the *index space* lies on the subspace spanned by the gradient.

In contrast with the work of Dalalyan, Juditsky, and Spokoiny (2008) our approach is local and it is therefore possible to retrieve a different subspace in different regions of $\mathbb{R}^d$. As localizing the estimator increases its variance, we choose to only identify the dimensions of interest instead of estimating the full projection matrix. We introduce *Gradient Guided Trees* in algorithm 3 to exploit the local aspect of our estimator in order to direct the cuts in a random tree: at each step, cuts are drawn randomly with probability proportional to estimated mean absolute gradient in the cell. We use the Random Survival Forest algorithm of Ishwaran and Kogalur

Algorithm 3 Node Splitting for Gradient Guided Trees

Require: (X, Y): training set, `Node`: indexes of points in the node

1: $\nabla r(X_i) \leftarrow$ `estimated gradient at` X_i, $\forall i \in$ `Node using` eq. (4.7)
2: $\omega \leftarrow \sum_{i \in \texttt{Node}} |\nabla r(X_i)|$
3: $K \leftarrow$ `sample` $\sqrt{d}$ `dimensions in` $\{1, \ldots, d\}$ `with probabilities` $\propto \omega$
4: $k, c \leftarrow$ `best threshold` c `and dimension` k
5: **return** k, c

(2007) as the best algorithm, which like most random forest models work by aggregating random trees in order to reduce the variance. Each separate random tree works by recursively splitting the space along the axes as shown in fig. 4.3 by selecting the cut that minimizes some homogeneity criterion of the leaves such as the log-rank statistic in the case of survival analysis (Shimokawa, Kawasaki, and Miyaoka [2015]; Robins and Finkelstein [2000]). In order to reduce the computational burden as well as introduce randomness to decorrelates the individual trees, randomness is introduced not only in the sampling mechanism but the choice of potential cuts. We introduce the knowledge of the gradient in this last step where instead of choosing candidate splits completely at random we choose with a probability proportional to the gradient. Note that in order to reduce the number of computations needed it is possible to precompute the gradient on the whole training set once at the beginning and then reuse the gradients of each individual example inside the node themselves by averaging the precomputed gradients over the cell members. We demonstrate

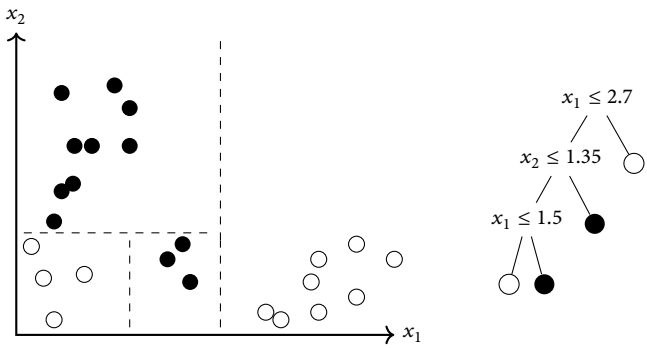


FIGURE 4.3: Recursive orthogonal splitting of the space in homogeneous cells by a tree.

	Description		Loss	
Dataset	n	d	RF	GGF
Wisconsin	569	30	$0.0352 \pm 3.29 \cdot 10^{-4}$	$\mathbf{0.0345} \pm 3.35 \cdot 10^{-4}$
Heart	303	13	$0.128 \pm 6.6 \cdot 10^{-4}$	$\mathbf{0.124} \pm 8.6 \cdot 10^{-4}$
Diamonds	53940	23	$680033 \pm 3.45 \cdot 10^{9}$	$\mathbf{664265} \pm 2.81 \cdot 10^{9}$
Gasoline	60	401	0.678 ± 0.451	$\mathbf{0.512} \pm 0.347$
SDSS	10000	8	$0.872 \cdot 10^{-3} \pm 4.50 \cdot 10^{-6}$	$\mathbf{0.776} \cdot 10^{-3} \pm 6.00 \cdot 10^{-6}$

TABLE 4.1: Performance of the two random forest algorithms on a 50-folds cross validation.

the improvements brought by guiding the cuts by the local information provided by the gradient by comparing the performance of a vanilla regression random forest with the same procedure but with local gradient information. We consider five datasets: the Breast Cancer Wisconsin (Diagnostic) Data Set introduced in Street, Wolberg, and Mangasarian (1993); the Heart Disease dataset introduced by Detrano et al. (1989); the classic Diamonds Price dataset; the Gasoline NIR dataset introduced by Kalivas (1997) and the Sloan Digital Sky Survey DR14 dataset of Abolfathi et al. (2018). We measure the L^2 loss by cross validation across 50 folds using the same hyperparameters for the growing of the forest in both the standard and gradient guided variants.

We denote by random forest (RF), forests grown from the standard classification and regression tree (CART) algorithm of Breiman et al. (1984) while gradient guided forest (GGF) denote forests grown from the *Gradient Guided Trees* previously introduced. As seen in table 4.1, gradient guided split sampling consistently outperform the vanilla variant. When all variables are relevant, as is the case when the variables were carefully selected by the practitioner with prior knowledge, our variant performs similarly to the original algorithm while performance is greatly improved when only a few variables are relevant, such as in the NIR dataset (Portier and Delyon [2014]).

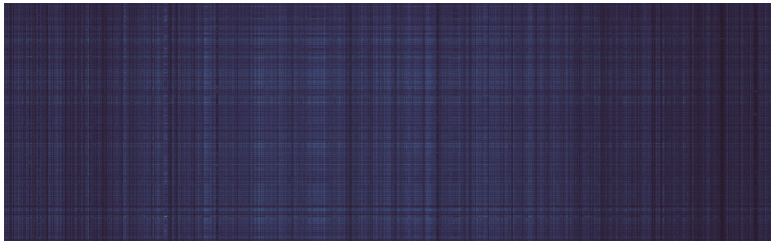


FIGURE 4.4: Map of the expression of 19946 genes from 1079 individuals

4.5.2 Survival Gradients

While the method presented here makes it possible to compute the gradient of a regression function in the standard setting, we are interested more specifically in this manuscript in the case of censored data. Luckily, the formulation used for our estimate of the gradient is a penalized ERM problem, and we have seen how to deal with censored observations in the ERM case in chapter 2 but reweighting the individual observations by their inverse probability of censoring. The local linear problem of eq. (4.7) then becomes

$$\underset{(r,\beta)\in\mathbb{R}^{d+1}}{\operatorname{argmin}} \sum_{i\in\hat{i}_k(x)} \frac{\delta_i}{\hat{S}_C(T_i- \mid X_i)} \left(T_i - r - \beta^\intercal (X_i - x)\right)^2 + \lambda\|\beta\|_1, \qquad (4.11)$$

in the right censored setting. Given the previous problem it is then possible to estimate the *survival gradient* of the regression function r and therefore select the dimensions of interest for the prediction of the survival.

We apply eq. (4.11) to the TCGA-BReast CAncer gene (BRCA) dataset which consists of gene expression quantification gathered by RNA sequencing.[123] The dataset consists of 19946 variables, after discarding 3 always constant variables, representing the *expression* of genes encoding various proteins, some suspected of playing a role in the susceptibility to cancer (see Fontanilla Ramirez [2020], for an example on Lamin B1). The TCGA-BRCA dataset in particular consists of patients diagnosed with breast cancer and focuses on the goal of identifying the various genetic risk factors of which the BRCAs genes have been shown to be examples. In our case we use normalized expression data of fig. 4.4 as our covariate of interest X and the time until death from the diagnostic as Y, approximately 80% of the observations are censored. While the map of the genome of fig. 4.4 is fairly uninformative and uniform, which is to be expected as its normalized in order to be readable, computing the gradient of the regression function paints a different picture. By using the same principle as earlier, we estimate the mean gradient of the dataset $\bar{\nabla} r$ by means of eq. (4.11). Given the very high number of dimensions and therefore the very high cost of distance computations, the use of an acceleration stucture

123: The data recorded corresponds to counts of deoxyribonucleic acid (DNA) sequences reverse transcribed from RNA sequences.

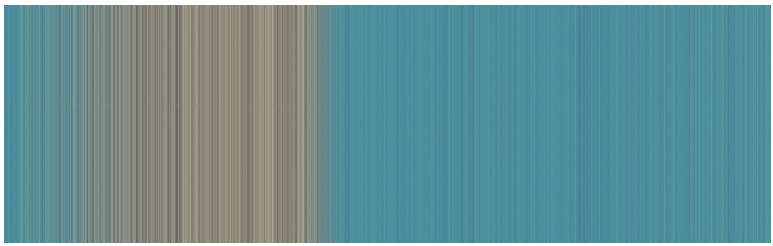


FIGURE 4.5: Gradient of the survival regression function.

is here primordial. For this specific example, instead of k-dimensional (KD)-trees, which results in exact nearest neighbour search, we use random tree projections as an approximate generalization of KD-trees as proposed by Hyvonen et al. (2016) and implemented in the `Shrike.jl` package. Patterns of variable importance clearly emerge in fig. 4.5 and can help direct attention to specific genes of interest.

4.5.3 Gradient Free Optimization

Many of the recent advances in the field of machine learning have been made possible in one way or another by advances in optimization; both in how well we are able to optimize complex function and what type of functions we are able to optimize if only locally. Recent advances in automatic differentiation as well as advances that push the notion of *what* can be differentiated have given rise to the notion of *differentiable programming* (see Innes et al. [2019]) in which a significant body of work can be expressed as the solution to a minimization problem usually then solved by gradient descent.

We study here the use of the local linear estimator of the gradient in algorithm 4 in cases where analytic or automatic differentiation is impossible, and compare it to a standard gradient free optimization technique as well as the oracle where the true gradient is known. While line 1 bears resemblance with Gaussian smoothing and could therefore be seen as analogous to gradient estimation via Gaussian smoothing (see Berahas et al. (2020)), two key differences here are the subsequent local linear step as well as the fact that the samples from line line 1 are not necessarily the samples used in the local linear estimator of line line 4.

We first minimize the standard but challenging Rosenbrock function for different values of d. which is defined as

$$f(x) = 100 \sum_{i=1}^{d-1} (x_{i+1} - x_i)^2 + (x_i - 1)^2.$$

We compare for reference our approach to the Nelder-Mead (simplex search) algorithm; a standard gradient free optimization technique. It

Algorithm 4 Estimated Gradient Descent

Require: x_0: initial guess, f: function $\mathbb{R}^d \mapsto \mathbb{R}$, M: budget
1: $X \leftarrow X_1, \ldots, X_M$ `with` $X_i \sim \mathcal{N}(x_0, \varepsilon \times I_d)$
2: $Y \leftarrow f(X) \overset{\text{def}}{=} f(X_1), \ldots, f(X_M)$
3: **while** `not StoppingCondition` **do**
4: $r, \delta \leftarrow$ `estimated gradient at` x `w.r.t` X, Y `using` eq. (4.7)
5: $X \leftarrow X, X_1, \ldots, X_M$ `with` $X_i \sim \mathcal{N}(\texttt{GradientStep}(x, \delta), \varepsilon \times I_d)$
6: $Y \leftarrow f(X)$
7: $x \leftarrow \operatorname{argmin}_{X_i}\{f(X_i)\}$
8: **end while**
9: **return** x

is apparent in fig. 4.6[124] that estimating the gradient yields a significant advantage compared to traditional gradient-free techniques that usually have to rely on bounding arguments and feasible regions and therefore scale unfavourably with the dimension. As our approach uses a nearest neighbours formulation for the gradient estimate, we are able to efficiently reuse past samples in the current estimate of the gradient; this makes it possible to achieve a sufficiently accurate estimate of the gradient even in high dimensions. We compare in fig. 4.7 the approach developed previously to the estimators proposed by Yining Wang et al. (2018) and Fan (1992). As the approach proposed by Yining Wang et al. (2018) includes the use of *mirror descent*, for fairness, we have implemented our proposed gradient descent algorithm of algorithm 4 using our estimator as well as those of Yining Wang et al. (2018) and Fan (1993) for the gradient where we permit the reuse of previous samples where appropriate. We then reimplemented the mirror descent algorithm of Yining Wang et al. (2018) with the previous estimators of the gradient. We observe in fig. 4.7 that our method compares favourably: our estimator is able to reuse past samples in its gradient estimation and has therefore access to a better gradient estimate for a fixed, given number of function evaluations. We apply the previous method to the minimization of the log-likelihood of a logistic model on the UCI's Adult data set, consisting of 48842 observations and 14 attributes represented as $\theta \in \mathbb{R}^{101}$ once one-hot encoded and an intercept added.

$$\mathcal{L}_\theta(X) = -\sum_i Y_i \log\left(1 + \exp(-\theta^\intercal X_i)\right) - (1 - Y_i)\log\left(1 + \exp(\theta^\intercal X_i)\right).$$

We also compare the effective CPU wall time needed to reach a given log-likelihood in order to give a more comprehensive view of the relative performance of the multiple algorithms. Given that the time per iteration can vary greatly depending on the cost of evaluations and the cost of the

124: The number of function evaluations does not have any meaning for the true gradient. We use here that 1 estimated gradient step $\approx$ 50 function evaluations. 5000 function evaluations therefore equate to 100 gradient steps.

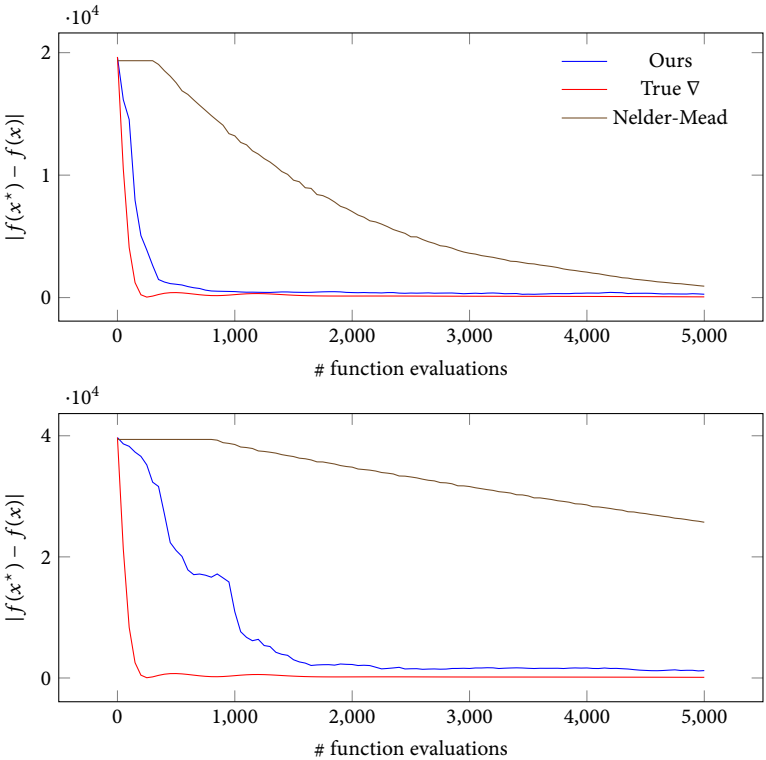


FIGURE 4.6: Nesterov Gradient Descent on the rosenbrock function for $d = 50$ (top) and $d = 100$ (bottom) w.r.t. the number of evaluations.

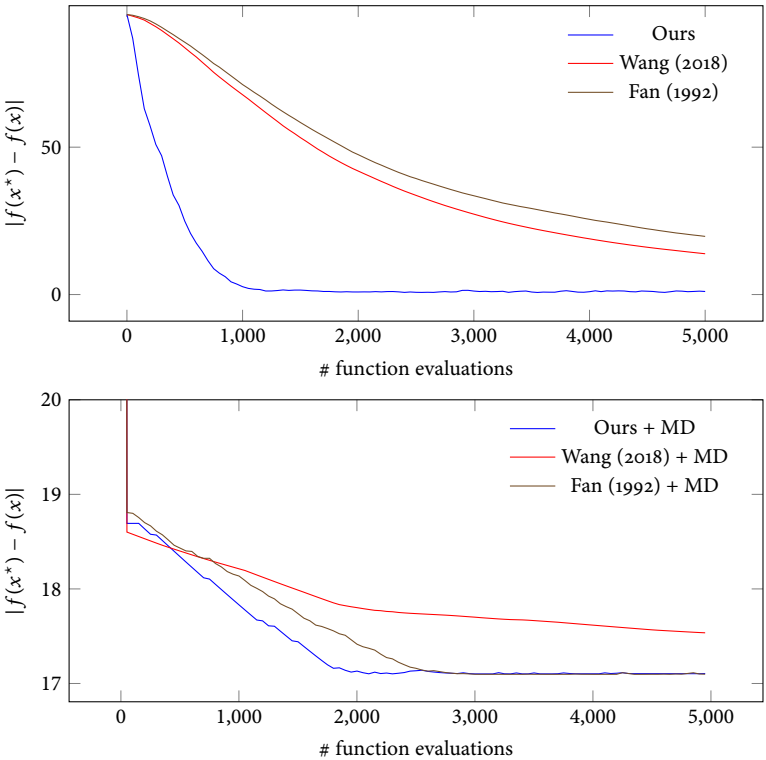


FIGURE 4.7: Nesterov Gradient Descent (top) and Mirror Gradient Descent (bottom) on the Rosenbrock function for $d = 100$.

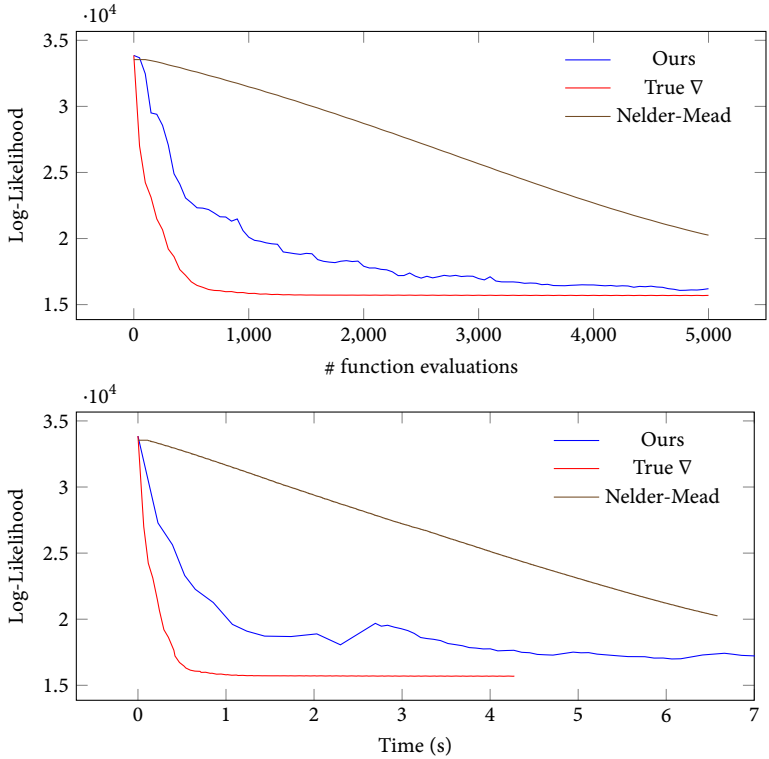


FIGURE 4.8: Log-likelihood of the logistic regression on a test set, trained by Nesterov Gradient Descent with respect to the number of evaluations (top) and time (bottom).

gradient procedures, it is important to use both the number of evaluations and the time metric jointly with the former being more relevant as the cost of individual function evaluations increases.

4.5.4 Disentanglement

Disentangled representation learning aims to learn a representation of the input space such that the independent dimensions of the representation each encode separate but meaningful attributes of the original feature space. If the space of interest is the space of *faces*, a disentangled representation would then for example be a lower-dimensional space where one dimension encodes the sex of the subject, another its age, and so forth. We show here how our estimator can be useful for retrieving the dimensions associated with a concept in a supervised manner.

A β-VAE (Higgins et al. [2017]) model is trained on the `CACD2000` dataset of celebrity faces with age labels to first build low-dimensional a representation of the images and then extract the direction relating to age. We

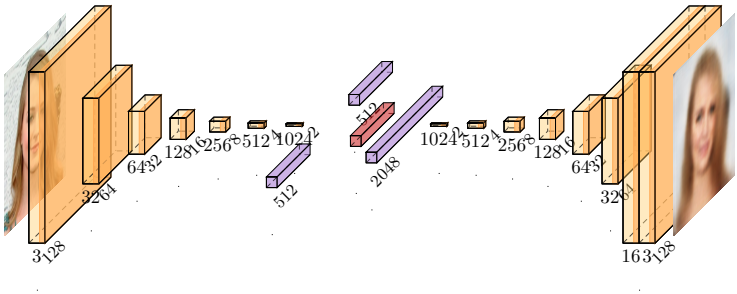


FIGURE 4.9: Encoder-Decoder Architecture used for this work

learn $\mathcal{E}_\Phi$ and $\mathcal{D}_\theta$ parameterizing q_Φ and p_θ, to minimize the loss

$$\mathcal{L}(\theta, \Phi; x, z, \beta) = \mathbb{E}_{q_\Phi(z|x)}\left[\log p_\theta(x \mid z)\right] - \beta D_{KL}(q_\Phi x \parallel x), \qquad (4.12)$$

where β acts as a constraint on the representational power of the latent distribution; $\beta = 1$ leads to the standard VAE formulation of Kingma and Welling (2014) while $\beta > 1$ increases the level of disentanglement. We use a standard symmetrical encoder-decoder architecture for the variational autoencoder, schematically presented in Figure fig. 4.9. All the relevant implementation details can be found in the `Julia` code in the `git` repository provided earlier.[125] We learn a 512-dimensional representation of the 128×128 images and encode all the `CACD2000` images. Once all the images have been encoded in $\mathbb{R}^{512}$ it is possible to use the local linear estimator of the gradient studied in this work to derive the gradient of the age with respect to the latent variable, making it possible to produce a new version of the input image that appears either older or younger as done in fig. 4.10. By computing a local estimate of the gradient, we are able to derive a more meaningful change when the age is not perfectly disentangled.

125: See `git.sr.ht/~aussetg/locallinear`.

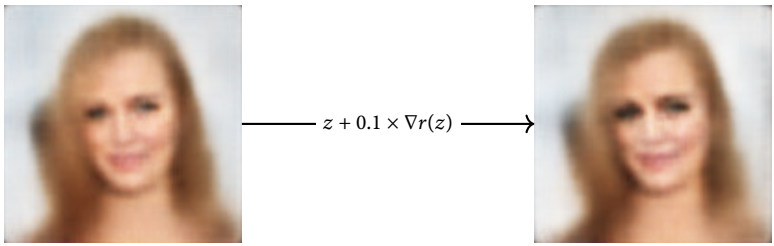


FIGURE 4.10: Extracting the direction of interest for aging.

Note that the quality of the image reconstruction and generation is here solely limited by the choice of the encoding and decoding model and is not related to the methods introduced in this chapter, significant advances in the quality of the decoding have been made in the recent years and if a

better quality and less blurry decoded output is desired we encourage the reader to replace the decoder with a `PixelCNN` architecture such as presented in Salimans et al. (2017). The quality of the gradient is also significantly impacted by the quality of the annotations as `CACD200` is an automatically annotated and noisy dataset.

Using our estimator it is possible to estimate the gradient ∇r of $r(z) = \mathbb{E}[Y \mid Z = z]$ with respect to the latent variable Z, as illustrated in fig. 4.10. It is then possible to analyse the sparsity of ∇r to quantify the quality of the disentanglement for varying level of β by quantifying how far from a single dimension the gradient for the age is concentrated. As the true dimension is unknown, we instead measure the angular distance to all dimensions reweighted by the magnitudes of the partial derivatives:

$$\sum_{k=1}^{d} \frac{|\bar{\nabla}_k r|}{|\hat{\nabla} r(x)|} \cos(e_i, |\bar{\nabla} r|)\,,$$

where $\cos(a, b)$ is the cosine similarity of a and b such that

$$\cos(a, b) = \frac{a \cdot b}{\|a\| \times \|b\|},$$

and measures the similarity between the direction of the vectors independently of their norm and where $\bar{\nabla} r$ is the mean gradient of the dataset, that is

$$|\bar{\nabla} r| = \frac{1}{n} \sum_{i=1}^{n} |\hat{\nabla} r(X_i)|\,.$$

We observe in fig. 4.11 that as β increases the age slowly become disentangled, as expected if one considers the age to be an important and independent characteristic of human faces.

While not an entirely adequate metric for disentanglement, not only because disentanglement does not necessarily require the dimensions to be the one an observer expected but more importantly because this metric requires an annotated dataset; we believe this metric can be useful for practitioners. By measuring how close the estimated gradients are to the axis, with respect to an annotated dataset of characteristics of interest, a practitioner can ensure his model is sufficiently disentangled for downstream tasks such as face manipulation by a user. We also believe it is possible to design an end-to-end differentiable framework in order to force disentanglement to consider the characteristics of interest: our estimator is the solution to a convex optimization program and as such admits an adjoint; it is therefore possible to fit a local linear estimator inside an automatic differentiation framework such as done in Agrawal et al. (2019).

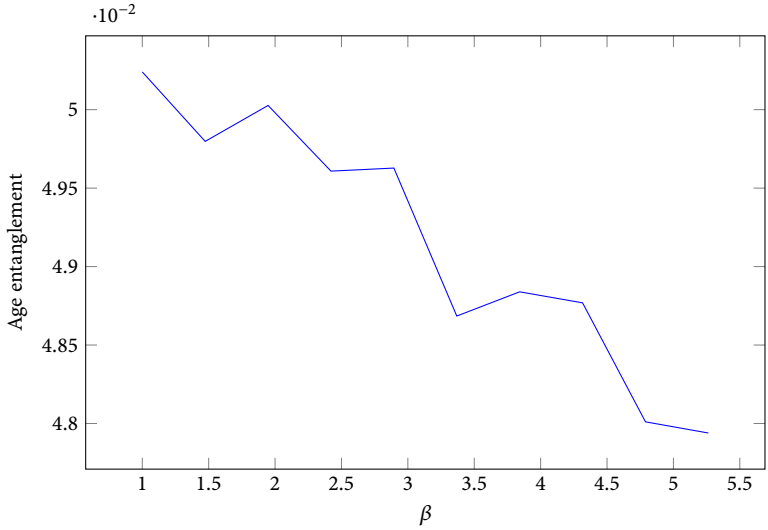


FIGURE 4.11: Quality of disentanglement with respect to the age

4.6 Conclusion

The *curse of dimensionality* has proven to be a thorn in the side of both theoricians as well as practictioners for as long as they have been analyzing data. It is however possible, as we have seen at the very beggining of this chapter in §4.2 to alleviate the problem, both in practice and in the bounds of the various results of the machine learning, by thinking about the geometry of the problem at hand in order to reduce the raw dimension of the data to a more manageable intrinsic dimension. In this chapter we have adopted in part this *geometry of the data* point of view by taking a supervised approach to the problem. While many dimension reduction methods model the data as lying on some manifold $\mathcal{M}$ of lower dimension than the overall space $\mathbb{X}$, we chose here to study the geometry of the tangent space of the regression function, that is the space spanned by the gradient of the regressor. By assuming that the gradient of the regression function is sparse, we are able to build a sparse estimator of the gradient, which then itself yields a lower dimensional subspace which we take as our effective dimension reduction (EDR) space. While the results presented here are a stark improvement to those of Jiang (2019), as the bound in Theorem 4.2 involves the degree of sparsity of the gradient. The quality of the EDR space obtained by our approach is validated experimentally both in §4.5.1 and §4.5.2 where we also show the usefulness of local methods compared to the usual global methods. Of course, the usefulness of the ability to estimate accurately the gradient of a function doesn't stop at finding priviledged dimensions as we have shown in §4.5.3 as well as

in a lesser measure in §4.5.4, therefore any further improvements that better exploit the characteristics of the gradient are of great interest. Note that the bound in Theorem 4.2, while taking into account the sparsity, still involves the dimension d of the full space which seems at odds with our goal of ignoring the irrelevant dimensions in order to free ourselves from the curse of dimensionality. It should be possible however to obtain similar bounds with d replaced by m the true EDR dimension. Dalalyan, Juditsky, and Spokoiny (2008) manages to obtain such bounds in the specific case of multi-index regression but similar bounds also exist in the more general *manifold* setting as shown in Aswani, Bickel, and Tomlin (2011), it is therefore natural to seek in future works to try bridging the gap between those results.

As banks such as BNP Paribas have access to virtually all the financial data of their internal clients, the number of dimensions available for prediction is naturally large even before any augmentation. After adding third-party data, either from external data providers, or by adding the unstructured data acquired from reports or news, the amount of data can quickly become unwieldy for most methods. Selecting the most important variables is therefore a necessary step in order to be able to make use of this data inside models more complex than generalized linear models (GLMs),[126] without resorting to the extreme of keeping the rating as unique variable. Ratings aggregate many completely different companies under a single general umbrella and do not offer a sufficiently granular view of the data. By incorporating many more variables, and therefore dimensions, it should be possible to build *individual* estimation of the credit risk of each company, hopefully leading to the construction of better portfolios.

126: GLMs are in practice incredibly powerful and therefore represent a very good baseline. Their relative simplicity and popularity, however, mean that they enjoy incredibly fast and optimized solvers making their use incredibly large data possible.

4.7 Proofs

We start by proving auxiliary results necessary for the intermediary proofs in §4.7.2 as well the main theorem in §4.7.3.

4.7.1 Auxiliary Results

As a first go, we recall or prove various auxiliary results that are involved in the proof of Theorem 4.1, and in that of Theorem 4.2 as well. The following inequality follows from the well-known Chernoff bound, (see e.g. Boucheron, Lugosi, and Massart [2013]).

Lemma 4.3. *Let $(Z_i)_{i\geq 1}$ be a sequence of i.i.d. random variables valued in $\{0,1\}$. Set $\mu \overset{\text{def}}{=} n\mathbb{E}[Z_1]$ and $S \overset{\text{def}}{=} \sum_{i=1}^{n} Z_i$. For any $\delta \in (0,1)$ and all $n \geq 1$,*

we have with probability $1 - \delta$*:*

$$S \geq \left(1 - \sqrt{\frac{2\log(1/\delta)}{\mu}}\right)\mu.$$

In addition, for any $\delta \in (0,1)$ *and* $n \geq 1$*, we have with probability* $1 - \delta$*:*

$$S \leq \left(1 + \sqrt{\frac{3\log(1/\delta)}{\mu}}\right)\mu.$$

Proof. Using the Chernoff lower tail (Boucheron, Lugosi, and Massart [2013]), for any $t > 0$ and $n \geq 1$, it holds that

$$\mathbb{P}\left(S < (1-t)\mu\right) \leq \left(\frac{\exp(-t)}{(1-t)^{1-t}}\right)^{\mu}.$$

Because for any $t \in (0,1)$

$$\frac{\exp(-t)}{(1-t)^{1-t}} \leq \exp\left(-t^2/2\right),$$

we obtain that for any $t > 0$ and $n \geq 1$,

$$\mathbb{P}\left(S < (1-t)\mu\right) \leq \exp\left(-\frac{t^2\mu}{2}\right),$$

the bound being obvious when $t \geq 1$. In the previous bound, choose $t = \sqrt{2\log(1/\delta)/\mu}$ to get the stated inequality. The second inequality is obtained by inverting the Chernoff upper tail:

$$\mathbb{P}\left(S > (1+t)\mu\right) \leq \left(\frac{\exp(t)}{(1+t)^{1+t}}\right)^{\mu}.$$

□

The following inequality is a well-known concentration inequality for sub-Gaussian random variables (see e.g. Boucheron, Lugosi, and Massart [2013]).

Lemma 4.4. *Suppose that* Z *is sub-Gaussian with parameter* $s^2 > 0$*, i.e.* Z *is real-valued, centred and for all* $\lambda > 0$

$$\mathbb{E}\left[\exp(\lambda Z)\right] \leq \mathbb{E}\left[\exp\left(\lambda^2 s^2/2\right)\right],$$

then with probability $1 - \delta$*,*

$$|Z| \leq \sqrt{2s^2\log(2/\delta)}.$$

We shall also need a concentration inequality tailored to VC classes of functions. The result stated in Lemma 4.5 below is mainly a consequence of the work of Giné and Guillou (2001). Our formulation is slightly different, the role played by the VC constants (v and A below) being clearly quantified.

Let $\mathcal{F}$ be a bounded class of measurable functions defined on $\mathcal{X}$. Let U be a uniform bound for the class $\mathcal{F}$, i.e. $|f(x)| \leq U$ for all $f \in \mathcal{F}$ and $x \in \mathcal{X}$. The class $\mathcal{F}$ is called VC with parameters (v, A) and uniform bound U if

$$\sup_Q \mathcal{N}(\varepsilon U, \mathcal{F}, L_2(Q)) \leq \left(\frac{A}{\varepsilon}\right)^v,$$

where $\mathcal{N}(., \mathcal{F}, L_2(Q))$ denotes the covering numbers of the class $\mathcal{F}$ relative to $L_2(Q)$ (see e.g. van der Vaart and Wellner [1996]). For notational simplicity and with no loss of generality, we require in the definition of a VC class that $A \geq 3\sqrt{e}$ and $v \geq 1$. We take $\sigma^2 \geq \sup_{f\in\mathcal{F}} \mathbb{V}[f(X_1)]$ and work under the condition

$$\sqrt{n}\sigma \geq c_1 \sqrt{U^2 v \log\left(\frac{AU}{\sigma\delta}\right)}, \tag{4.13}$$

where the constants c_1 and c_2 are specified in the following statement.

Lemma 4.5. *Let $\mathcal{F}$ be a VC class of functions with parameters (v, A) and uniform bound $U > 0$ such that $\sigma \leq U$. Let $n \geq 1$ and $\delta \in (0, 1)$. There are two positive universal constants c_1 and c_2 such that, under the condition of eq. (4.13), we have with probability $1 - \delta$,*

$$\sup_{f\in\mathcal{F}} \left| \sum_{i=1}^n (f(X_i) - \mathbb{E}[f(X_1)]) \right| \leq c_2 \sqrt{n\sigma^2 v \log\left(\frac{AU}{\sigma\delta}\right)}.$$

Proof. Set $\lambda = v \log(AU/\sigma)$. Using Giné and Guillou (2001), equation (2.5) and (2.6), we get

$$\begin{aligned}
\mathbb{E}\left[\sup_{f\in\mathcal{F}} \left| \sum_{i=1}^n (f(X_i) - \mathbb{E}[f(X_1)]) \right| \right] &\leq C\sqrt{\lambda}\left(\sqrt{n}\sigma + U\sqrt{\lambda}\right) \\
&\leq 2C\sqrt{n\sigma^2\lambda}, \\
\mathbb{E}\left[\sup_{f\in\mathcal{F}} \left| \sum_{i=1}^n (f(X_i) - \mathbb{E}f(X_1))^2 \right| \right] &\leq \left(\sqrt{n}\sigma + KU\sqrt{\lambda}\right)^2 \\
&\leq 4n\sigma^2 \stackrel{\text{def}}{=} V,
\end{aligned}$$

where $C > 0$ and $K > 0$ are two universal constants. Both previous inequalities are obtained by taking c_1 large enough. Let

$$Z = \sup_{f\in\mathcal{F}} \left| \sum_{i=1}^n (f(X_i) - \mathbb{E}[f(X_1)]) \right|,$$

We recall Talagrand's inequality (Talagrand [1996], Theorem 1.4; or Giné and Guillou [2001], eq. (2.7)), for all $t > 0$,

$$\mathbb{P}\left(|Z - \mathbb{E}[Z]| > t\right) \leq K' \exp\left(-\frac{t}{2K'U} \log\left(1 + 2t\frac{U}{V}\right)\right),$$

where $K' > 1$ is a universal constant. Using the fact that for all $t \geq 0$

$$\frac{t}{2 + 2t/3} \leq \log(1 + t),$$

we get

$$\mathbb{P}\left(|Z - \mathbb{E}[Z]| > t\right) \leq K' \exp\left(-\frac{t^2}{2K'(V + 2tU/3)}\right).$$

Inverting the bound, we find that for any $\delta \in (0, 1)$, with probability $1 - \delta$,

$$\begin{aligned}|Z - \mathbb{E}[Z]| &\leq \sqrt{2K'V \log\left(\frac{K'}{\delta}\right)} + \frac{4}{3}K'U \log\left(\frac{K'}{\delta}\right) \\ &\leq \sqrt{2K'VK'' \log\left(\frac{2}{\delta}\right)} + \frac{4}{3}K'UK'' \log\left(\frac{2}{\delta}\right),\end{aligned}$$

for some $K'' > 0$. Taking c_1 large enough and using that $AU/\sigma > 2$, we ensure that

$$2V = 8n\sigma^2 \geq \left(\frac{4U}{3}\right)^2 K'K'' \log\left(\frac{2}{\delta}\right).$$

Then using the previous bound on the expectation, it follows that with probability $1 - \delta$,

$$\begin{aligned}|Z| &\leq 2C\sqrt{n\sigma^2\lambda} + 2\sqrt{8n\sigma^2K'K'' \log(2/\delta)} \\ &= 2C\sqrt{n\sigma^2}\left(\sqrt{\lambda} + \sqrt{8K'K'' \log(2/\delta)}\right).\end{aligned}$$

We then conclude by using the bound $\sqrt{a} + \sqrt{b} \leq \sqrt{2}\sqrt{a + b}$. □

4.7.2 Intermediary Results

We now prove some intermediary results used in the core of the proof of the main results. We first define

$$\overline{\tau}_k = \left(\frac{2k}{nb_f V_d}\right)^{1/d}.$$

Proposition 4.6. *Suppose that Assumption 4.1 is fulfilled and that* $\overline{\tau}_k \le \tau_0$. *Then for any* $\delta \in (0,1)$ *such that* $k \ge 4\log(n/\delta)$, *we have with probability* $1-\delta$:

$$\hat{\tau}_k(x) \le \overline{\tau}_k.$$

Proof. Using Assumption 4.1 yields

$$\begin{aligned}
\mathbb{P}(X \in \mathcal{B}(x, \overline{\tau}_k)) &= \int_{\mathcal{B}(x,\overline{\tau}_k)} g \\
&\ge b_f \int_{\mathcal{B}(x,\overline{\tau}_k)} \mathrm{d}\lambda \\
&= b_f V_d \overline{\tau}_k^d \\
&= \frac{2k}{n}.
\end{aligned}$$

Consider the set formed by the n balls $\mathcal{B}(x, \overline{\tau}_k), 1 \le k \le n$. From Lemma 4.3 with $Z_i = \mathbb{1}_{\mathcal{B}(x,\overline{\tau}_k)}(X_i)$, $\mu \ge 2k$, and the union bound, we obtain that for all $\delta \in (0,1)$ and any $k = 1, \ldots, n$:

$$\sum_{i=1}^{n} \mathbb{1}_{\mathcal{B}(x,\overline{\tau}_k)}(X_i) \ge \left(1 - \sqrt{\frac{2\log(n/\delta)}{2k}}\right) 2k.$$

As $k \ge 4\log(n/\delta)$, it follows that

$$\begin{aligned}
\sum_{i=1}^{n} \mathbb{1}_{\mathcal{B}(x,\overline{\tau}_k)}(X_i) &\ge k - \left(\sqrt{4k\log\left(\frac{n}{\delta}\right)} - k\right) \\
&\ge k.
\end{aligned}$$

Hence $\mathbb{P}_n(\mathcal{B}(x, \overline{\tau}_k)) \ge k/n$, denoting by $\mathbb{P}_n$ the empirical distribution of the X_i's. By definition of $\hat{\tau}_k(x)$ it holds that $\hat{\tau}_k(x) \le \overline{\tau}_k(x)$. □

Define

$$\underline{\tau}_k = \left(\frac{k}{2nU_f V_d}\right)^{1/d},$$

and its neighbourhood

$$\underline{i}_k(x) = \left\{j : X_j \in \mathcal{B}(x, \underline{\tau}_k(x))\right\}.$$

Proposition 4.7. *Suppose that Assumption 4.1 is fulfilled and that* $\underline{\tau}_k \le \tau_0$. *Then for any* $\delta \in (0,1)$ *such that* $k \ge 4\log(n/\delta)$, *we have with probability* $1-\delta$:

$$\hat{\tau}_k \ge \underline{\tau}_k.$$

Proof. Using Assumption 4.1 yields

$$\begin{aligned}\mathbb{P}(X \in \mathcal{B}(x, \underline{\tau}_k)) &= \int_{\mathcal{B}(x,\underline{\tau}_k)} g \\ &\leq U_f \int_{\mathcal{B}(x,\underline{\tau}_k)} \mathrm{d}\lambda \\ &= U_f V_d \underline{\tau}_k^d \\ &= \frac{k}{2n}.\end{aligned}$$

Consider the set formed by the n balls $\mathcal{B}(x, \underline{\tau}_k)$, $1 \leq k \leq n$. From Lemma 4.3 with $Z_i = \mathbb{1}_{\mathcal{B}(x,\underline{\tau}_k)}(X_i)$, $\mu \leq k/2$, and the union bound, we obtain that for all $\delta \in (0,1)$ and $k = 1, \ldots, n$

$$\sum_{i=1}^{n} \mathbb{1}_{\mathcal{B}(x,\underline{\tau}_k)}(X_i) \leq \frac{k}{2}\left(1 + \sqrt{\frac{6}{k}\log\left(\frac{n}{\delta}\right)}\right).$$

Using that $k \geq 6\log(n/\delta)$, it follows that

$$\begin{aligned}\sum_{i=1}^{n} \mathbb{1}_{\mathcal{B}(x,\underline{\tau}_k)}(X_i) &\leq k + \sqrt{\frac{6}{4}k\log\left(\frac{n}{\delta}\right)} - \frac{k}{2} \\ &\leq k.\end{aligned}$$

Hence $\mathbb{P}_n(\mathcal{B}(x, \underline{\tau}_k)) \leq k/n$. By definition of $\hat{\tau}_n(k)(x)$ it holds that $\underline{\tau}_k \leq \hat{\tau}_k(x)$. □

Proposition 4.8. *Suppose that Assumption 4.2 is fulfilled. Then for any $\delta \in (0,1)$, we have with probability $1-\delta$:*

$$\left|\sum_{i \in \hat{\imath}_k(x)} \xi_i\right| \leq \sqrt{2k\sigma^2 \log\left(\frac{2}{\delta}\right)}.$$

where

$$\hat{\imath}_k(x) = \left\{j : X_j \in \mathcal{B}(x, \hat{\tau}_k(x))\right\}.$$

Proof. Set $w_i = \mathbb{1}_{\mathcal{B}(x,\hat{\tau}_k(x))}(X_i)$. Note that $\sum_{i=1}^n w_i^2 = k$ almost surely. The result follows from the application of Lemma 4.4 to the random variable $\sum_{i=1}^n \xi_i w_i$, which is sub-Gaussian with parameter $k\sigma^2$. To check this, it is

enough to write

$$\begin{aligned}\mathbb{E}\left[\exp\left(\lambda\sum_{i=1}^{n}\xi_i w_i\right)\right] &= \mathbb{E}\left[\mathbb{E}\left[\exp\left(\lambda\sum_{i=1}^{n}\xi_i w_i\right)\,\middle|\, X_1,\dots X_n\right]\right]\\ &= \mathbb{E}\left[\prod_{i=1}^{n}\mathbb{E}\left[\exp(\lambda\xi_i w_i)\mid X_1,\dots X_n\right]\right]\\ &\le \mathbb{E}\left[\prod_{i=1}^{n}\mathbb{E}\left[\exp(\lambda^2\sigma^2 w_i^2/2)\mid X_1,\dots X_n\right]\right]\\ &= \mathbb{E}\left[\exp\left(\lambda^2\sigma^2\sum_{i=1}^{n}w_i^2/2\right)\right]\\ &= \exp\left(\lambda^2\sigma^2 k/2\right).\end{aligned}$$

□

Proposition 4.9. *Suppose that Assumption 4.1 and Assumption 4.2 are fulfilled and that $\overline{\tau}_k \le \tau_0$. Let $\hat{H}_i \overset{\text{def}}{=} h_i(X_1,\dots,X_n)$ such that*

$$a_k = \sup_{i\in\bar{\imath}_k}|\hat{H}_i|\,.$$

Then for any $\delta \in (0,1)$ such that $k \ge 4\log(2n/\delta)$, we have with probability $1-\delta$:

$$\left|\sum_{i\in\hat{\imath}_k(x)}\xi_i\hat{h}_i\right| \le \sqrt{2k\sigma^2 a_k^2\log\left(\frac{4}{\delta}\right)}.$$

Proof. Set $w_i = \mathbb{1}_{\mathcal{B}(x,\hat{\tau}_k(x))}(X_i)$. Note that $\sum_{i=1}^n w_i^2 = k$ almost surely. The result follows from the fact that conditioned upon $X_1,\dots,X_n$, the random variable $\sum_{i=1}^n \xi_i h_i w_i$ is sub-Gaussian with parameter $\sigma^2 k\hat{a}_k^2$ with

$$\hat{a}_k = \sup_{i\in\bar{\imath}_k}|\hat{H}_i|\,.$$

To check this, it suffices to write

$$\begin{aligned}\mathbb{E}\left[\exp\left(\lambda\sum_{i=1}^{n}\xi_i\hat{H}_i w_i\right)\mid X_1,\dots X_n\right] &= \prod_{i=1}^{n}\mathbb{E}\left[\exp(\lambda\xi_i\hat{H}_i w_i)\mid X_1,\dots X_n\right]\\ &\le \prod_{i=1}^{n}\exp(\lambda^2\sigma^2\hat{H}_i^2 w_i^2/2)\\ &= \exp\left(\lambda^2\sigma^2\sum_{i=1}^{n}\hat{H}_i^2 w_i/2\right)\\ &\le \exp(\lambda^2\sigma^2 k\hat{a}_k^2/2)\,.\end{aligned}$$

Then, for any $t > 0$,

$$\begin{aligned}
\mathbb{P}\left(\left|\sum_{i=1}^n \xi_i h_i w_i\right| > t\right) &\le \mathbb{P}\left(\left|\sum_{i=1}^n \xi_i h_i w_i\right| > t, \hat{\tau}_k(x) \le \tau_k(x)\right) \\
&\quad + \mathbb{P}\left(\hat{\tau}_k(x) \le \tau_k(x)\right) \\
&\le \mathbb{E}\left[\mathbb{P}\left(\left|\sum_{i=1}^n \xi_i h_i w_i\right| > t \mid X_1, \dots, X_n\right) \mathbb{1}_{\hat{\tau}_k(x) \le \tau_k(x)}\right] \\
&\quad + \mathbb{P}(\hat{\tau}_k(x) \le \tau_k(x)) \\
&\le \mathbb{E}\left[2\exp\left(-\frac{t^2}{2k\sigma^2 \hat{a}_k^2}\right) \mathbb{1}_{\hat{\tau}_k(x) \le \tau_k(x)}\right] \\
&\quad + \mathbb{P}\left(\hat{\tau}_k(x) \le \tau_k(x)\right) \\
&\le 2\exp\left(-\frac{t^2}{2k\sigma^2 a_k^2}\right) + \mathbb{P}\left(\hat{\tau}_k(x) \le \tau_k(x)\right).
\end{aligned}$$

We obtain the result by choosing $t = \sqrt{2k\sigma^2 a_k^2 \log(4/\delta)}$ and applying Proposition 4.6 (to obtain that $\mathbb{P}(\hat{\tau}_k(x) \le \tau_k(x)) \le \delta/2$). □

Proposition 4.10. *Suppose that Assumption 4.1 and Assumption 4.4 are fulfilled. Let $\tau > 0$, $n \ge 1$, and $\delta \in (0,1)$ such that $\tau \le \tau_0$ and $24nU_f(2\tau)^d \ge \log(2d^2/\delta)$, then with probability $1-\delta$,*

$$\begin{aligned}
\max_{1 \le j,j' \le d} \Bigg| \sum_{i=1}^n \Big(&(X_{i,j} - x)(X_{i,j'} - x)^\intercal \mathbb{1}_{\mathcal{B}(x,\tau)}(X_i) \\
&- \mathbb{E}\left[(X_{1,j} - x)(X_{1,j'} - x)^\intercal \mathbb{1}_{\mathcal{B}(x,\tau)}(X_1)\right] \Big) \Bigg| \\
&\le (2\tau)^2 \sqrt{\frac{2U_f n(2\tau)^d}{3} \log\left(\frac{2d^2}{\delta}\right)}.
\end{aligned}$$

Proof. We use Bernstein's inequality: for any collection $(Z_1, \dots, Z_n)$ of independent zero-mean random variables such that for all $i = 1, \dots, n$, $|Z_i| \le m$ and $\mathbb{E}Z_i^2 \le v$, it holds that with probability $1-\delta$,

$$\left|\sum_{i=1}^n Z_i\right| \le \sqrt{2nv \log\left(\frac{2}{\delta}\right)} + \frac{m}{3} \log\left(\frac{2}{\delta}\right).$$

Applying this with

$$\begin{aligned}
W_i &= \frac{(X_{i,j} - x)}{2\tau} \frac{(X_{i,j'} - x)}{2\tau} \mathbb{1}_{\mathcal{B}(0,\tau)}(X_i), \\
Z_i &= W_i - \mathbb{E}[W_i],
\end{aligned}$$

we can use

$$|Z_i| \leq 2\,|W_i| \leq 1/4 = m,$$

and

$$\begin{aligned}
\mathbb{E}\left[(W_i - \mathbb{E}\left[W_i\right])^2\right] &\leq \mathbb{E}\left[W_i^2\right] \\
&= \mathbb{E}\left[\left|\frac{(X_{i,j} - x)}{2\tau}\frac{(X_{i,j'} - x)}{2\tau}\right|^2 \mathbb{1}_{\mathcal{B}(0,\tau)}(X_i)\right] \\
&= \int \left|\frac{(y_j - x)}{2\tau}\frac{(y_{j'} - x)}{2\tau}\right|^2 \mathbb{1}_{\mathcal{B}(0,\tau)}(y) f(y)\,\mathrm{d}y \\
&\leq U_f \int \left|\frac{(y_j - x)}{2\tau}\frac{(y_{j'} - x)}{2\tau}\right|^2 \mathbb{1}_{\mathcal{B}(0,\tau)}(y)\,\mathrm{d}y \\
&= U_f (2\tau)^d \int \left|u_j u_{j'}\right|^2 \mathbb{1}_{\mathcal{B}(0,1/2)}(u)\,\mathrm{d}u \\
&\leq U_f (2\tau)^d \int \frac{u_j^2 + u_{j'}^2}{2} \mathbb{1}_{\mathcal{B}(0,1/2)}(u)\,\mathrm{d}u \\
&= U_f (2\tau)^d \int u_1^2 \mathbb{1}_{\mathcal{B}(0,1/2)}(u)\,\mathrm{d}u \\
&= U_f (2\tau)^d \int_{-1/2}^{1/2} u_1^2\,\mathrm{d}u_1 \\
&= \frac{U_f (2\tau)^d}{12} \stackrel{\text{def}}{=} v.
\end{aligned}$$

We have shown that, with probability $1 - \delta$,

$$\left|\sum_{i=1}^n Z_i\right| \leq \sqrt{\frac{nU_f(2\tau)^d}{6}\log\left(\frac{2}{\delta}\right)} + \frac{1}{12}\log\left(\frac{2}{\delta}\right).$$

Because $24nU_f(2\tau)^d \geq \log(2/\delta)$, we obtain that

$$\left|\sum_{i=1}^n Z_i\right| \leq 2\sqrt{\frac{nU_f(2\tau)^d}{6}\log\left(\frac{2}{\delta}\right)}.$$

Replacing δ by δ/d^2 and using the union bound, we get the desired result. □

An important quantity in the framework we develop is

$$\sum_{i \in \hat{i}_k(x)} (X_{i,j} - x_j),$$

for which we provide an upper bound in the following theorem. Note that we improve upon the straightforward bound of $k\hat{\tau}_k(x)$ which is unfortunately not enough for the analysis carried out here. We shall work with the following assumption

$$C_1 \log\left(\frac{Dn}{\delta}\right) \le k \le C_2 n, \tag{4.14}$$

where the two constants $C_1 > 0$ and $C_2 > 0$ are given in the following proposition.

Proposition 4.11. *Suppose that Assumption 4.1 and Assumption 4.4 are fulfilled. Let $n \ge 1$, $k \ge 1$ and $\delta \in (0,1)$. There exist universal positive constants C_1, C_2, and C_3 such that, under eq. (4.14), we have with probability $1-\delta$,*

$$\max_{j=1,\dots,d} \left| \sum_{i \in \hat{i}_k(x)} \left(X_{i,j} - x_j\right) \right| \le C_3 \left(\overline{\tau}_k \sqrt{k \log\left(\frac{nd}{\delta}\right)} + \frac{Lk\overline{\tau}_k^2}{b_f} \right).$$

Proof. Taking C_1 greater than 4, we ensure that $k \ge 4\log(2n/\delta)$. Taking C_2 small enough, we guarantee that $\overline{\tau}_k \le \tau_0$. From Proposition 4.6, we have that $\hat{\tau}_k(x) \le \overline{\tau}_k$ is valid with probability $1 - \delta/2$. Let

$$\mu(\tau) = \mathbb{E}\left[(X_1 - x)\, \mathbb{1}_{\mathcal{B}(x,\tau)}(X_1) \right].$$

Consider the following decomposition

$$\begin{aligned}
\left| \sum_{i \in \hat{i}_k(x)} \left(X_{i,j} - x_j\right) \right| &\le \left| \sum_{i \in \hat{i}_k(x)} \left(\left(X_{i,j} - x_j\right) - \mu_j\left(\hat{\tau}_k(x)\right) \right) \right| + n\mu_j\left(\hat{\tau}_k(x)\right) \\
&\le \sup_{0 < \tau \le \overline{\tau}_k} \left| \sum_{i \in i_\tau(x)} \left(\left(X_{i,j} - x_j\right) - \mu_j(\tau) \right) \right| + n\mu_j\left(\hat{\tau}_k(x)\right).
\end{aligned}$$

where

$$i_\tau(x) = \left\{ j : X_j \in \mathcal{B}(x,\tau) \right\}.$$

Notice that

$$\begin{aligned}
\mu(\tau) &= \int (y - x) \mathbb{1}_{\mathcal{B}(x,\tau)}(y) f(y)\, \mathrm{d}y \\
&= (2\tau)^{1+d} \int_{\mathcal{B}(0,1/2)} v f(x + \tau v)\, \mathrm{d}v \\
&= (2\tau)^{1+d} \int_{\mathcal{B}(0,1/2)} v \left(f(x + \tau v) - f(x) \right) \mathrm{d}v.
\end{aligned}$$

Hence

$$\begin{aligned}\left|\mu_j(\tau)\right| &\leq \frac{L}{2}(2\tau)^{2+d}\int_{\mathcal{B}(0,1/2)} v_j \left\|v\right\|_\infty \mathrm{d}v \\ &\leq \frac{L}{8}(2\tau)^{2+d} \\ &= \frac{L}{8}(2\tau)^{2+d}.\end{aligned}$$

And we find

$$\begin{aligned}\sup_{j=1,\ldots,d}\left|\mu_j(\hat{\tau}_k)\right| &\leq \frac{L}{8}(2\overline{\tau}_k)^{2+d} \\ &= \frac{Lk}{b_f n}\overline{\tau}_k^2.\end{aligned}$$

The class of rectangles $\mathcal{R} = \{y \mapsto \mathbb{1}_{\mathcal{B}(x,\tau)}(y) : \tau > 0\}$ cannot shatter 2 points x_1 and x_2. Considering the case $\left\|x_1 - x\right\|_\infty < \left\|x_2 - x\right\|_\infty$, it fails to pick out x_2. Hence its VC index is $v = 2$. From Theorem 2.6.4 in van der Vaart and Wellner (1996), we have

$$\mathcal{N}\left(\varepsilon, \mathcal{R}, L_2(Q)\right) \leq Kv(4e)^v\left(\frac{1}{\varepsilon}\right)^{2(v-1)},$$

for any probability measure Q. Which therefore implies that

$$\mathcal{N}\left(\varepsilon, \mathcal{R}, L_2(Q)\right) \leq \left(A/\varepsilon\right)^2,$$

where A is a universal constant. As a result, the class

$$\mathcal{F}_j = \left\{y \mapsto \frac{(y - x_j)}{\overline{\tau}_k}\mathbb{1}_{\mathcal{B}(x,\tau)}(y) : \tau \in (0, \overline{\tau}_k]\right\},$$

which is uniformly bounded by 1, satisfies the exact same bound for its covering number, that is

$$\mathcal{N}\left(\varepsilon, \mathcal{F}_j, L_2(Q)\right) \leq \left(\frac{A}{\varepsilon}\right)^2.$$

We can therefore apply Lemma 4.5 with $v = 2$, A a universal constant, $U = 1$ and σ^2 defined as

$$\begin{aligned}\sigma^2 &\stackrel{\text{def}}{=} \mathbb{V}\left[\frac{(X_1 - x)_j}{\overline{\tau}_k}\mathbb{1}_{\mathcal{B}(x,\tau)}(X_1)\right] \\ &\leq \mathbb{E}[\mathbb{1}_{\mathcal{B}(x,\tau)}(X_1)] \\ &\leq \mathbb{E}[\mathbb{1}_{\mathcal{B}(x,\overline{\tau}_k)}(X_1)] \\ &\leq \frac{2kU_f}{nb_f} \\ &\leq \frac{4k}{n}.\end{aligned}$$

Equation (4.13) is valid under eq. (4.14) when C_1 (resp. C_2) is a large (resp. small) enough constant. The fact that $\sigma^2 \leq 1$ is provided by eq. (4.14) as well. We obtain that

$$\sup_{0<\tau\leq\overline{\tau}_k} \left| \sum_{i\in i_\tau(x)} \Big(\big(X_{i,j} - x_j\big) - \mu_j(\tau) \Big) \right| \leq \overline{\tau}_k C \sqrt{kd \log\left(\frac{n}{\delta}\right)},$$

where C is a universal constant (C should be large enough to absorb the other constants involved until now). Using the union bound, this bound is extended to a uniform bound over $j \in \{1, \ldots, d\}$. We then obtain the statement of the proposition. □

4.7.3 Proof of Theorem 4.2

We rely on the bias-variance decomposition expressed in eq. (4.9). On the first hand, we have

$$\begin{aligned} |r_k(x) - r(x)| &= \left| \frac{\sum_{i_k(x)} (r(X_i) - r(x))}{\sum_{i=1}^n \mathbb{1}_{\mathcal{B}(x,\hat{\tau}_k(x))}(X_i)} \right| \\ &\leq \sup_{y\in\mathcal{B}(x,\hat{\tau}_k(x))} |r(y) - r(x)| \\ &\leq L_1 \hat{\tau}_k(x). \end{aligned}$$

Applying Proposition 4.6 we obtain that, with probability $1 - \delta/2$,

$$|r_k(x) - r(x)| \leq L_1 \overline{\tau}_k.$$

On the other hand, we apply Proposition 4.8 to get that, with probability $1 - \delta/2$,

$$|\hat{r}_k(x) - r_k(x)| \leq \sqrt{\frac{2\sigma^2 \log(4/\delta)}{k}}.$$

4.7.4 Proof of Theorem 4.1

Denote by $\mathbb{X}$ the design matrix of the (local) regression problem

$$\begin{aligned} \mathbb{X} &= (X_{i_1}^\intercal, \ldots, X_{i_k}^\intercal)^\intercal, \\ \mathbb{Y} &= (y_{i_1}^\intercal, \ldots, y_{i_k}^\intercal)^\intercal. \end{aligned}$$

where for any $j = 1, \ldots, k$, i_j is such that $X_{i_j} \in \mathcal{B}(x; \hat{\tau}_k(x))$. Define

$$\begin{aligned} w &= \mathbb{Y} - \mathbb{X}\beta^*, \\ \hat{v} &= \hat{\beta}_k(x) - \beta^*. \end{aligned}$$

Following Hastie, Tibshirani, and Wainwright (2015), define

$$C(S,\alpha) = \left\{u \in \mathbb{R}^d : \left\|u_{\overline{S}}\right\|_1 \leq \alpha \left\|u_S\right\|_1\right\}.,$$

and let $\hat{\gamma}_n$ be defined as

$$\hat{\gamma}_n = \inf_{u \in C(S,3)} \frac{\|\mathbb{X}u\|_2^2}{k\|u\|_2^2}.$$

Hence, $\hat{\gamma}_n$ is the smallest eigenvalue (restricted to the cone) of the design matrix $\mathbb{X}$. From Lemma 11.1 in Hastie, Tibshirani, and Wainwright (2015), we have the following: whenever

$$\lambda \geq \frac{2}{k}\left\|\mathbb{X}^\intercal w\right\|_\infty,$$

it holds that

$$\hat{v} \in C(S,3),$$
$$\|\hat{v}\|_2 \leq 3\lambda \frac{\sqrt{|S_x|}}{\hat{\gamma}_n}.$$

Consequently, the proof will be completed if, with probability $1-\delta$,

$$\frac{2}{k}\left\|\mathbb{X}_j^\intercal w\right\|_\infty \leq \overline{\tau}_k\left(\sqrt{\frac{2\sigma^2\log(16d/\delta)}{k}} + L_2\overline{\tau}_k^2\right), \tag{4.15}$$

$$\hat{\gamma}_n \geq \frac{\overline{\tau}_k^2}{24\times 8}. \tag{4.16}$$

Proof of eq. (4.15). In the next few lines, we show that eq. (4.15) holds with probability $1-\delta/2$. By definition

$$\mathbb{X}^\intercal w = \sum_{i\in\hat{i}_k(x)} w_i^\intercal X_i^\intercal,$$

Using that $w_i = \xi_i + r(X_i) - \beta^{*\intercal}X_i$,

$$\begin{aligned}\mathbb{X}^\intercal w &= \sum_{i\in\hat{i}_k(x)} X_i^\intercal\xi_i + \sum_{i\in\hat{i}_k(x)} X_i^\intercal\left(r(X_i) - \beta^{*\intercal}X_i\right)\\ &= \sum_{i\in\hat{i}_k(x)} X_i^\intercal\xi_i + \sum_{i\in\hat{i}_k(x)} X_i^\intercal\left(r(X_i) - r(x) - \beta^{*\intercal}(X_i - x)\right),\end{aligned}$$

where we have used the covariance structure (with empirically centred terms) to derive the last line. Note that for any $\tau > 0$, $\max_{i:X_i\in\mathcal{B}(x,\tau)}\left|X_{i,j}^\intercal\right| \leq$

τ. Hence, from Proposition 4.9, because $\overline{\tau}_k \le \tau_0$ and $k \ge 4\log(8dn/\delta)$ (taking C_1 large enough), we have with probability $1 - \delta/(4d)$,

$$\left| \sum_{i \in \hat{i}_k(x)} X_{i,j}^\intercal \xi_i \right| \le \sqrt{2k\sigma^2 \overline{\tau}_k^2 \log\left(\frac{16d}{\delta}\right)}.$$

Moreover,

$$\begin{aligned} \sum_{i \in \hat{i}_k(x)} \left| X_{i,j}^\intercal \right| \times \left| r(X_i) - r(x) - g(x)^\intercal (X_i - x) \right| &\le k L_2 \hat{\tau}_k(x)^2 \max_{i \in \hat{i}_k(x)} \left| X_{i,j}^\intercal \right| \\ &\le k L_2 \hat{\tau}_k(x)^3. \end{aligned}$$

Using Proposition 4.6, because $k \ge 4\log(4dn/\delta)$, it holds, with probability $1 - \delta/(4d)$,

$$\sum_{i \in \hat{i}_k(x)} \left| X_{i,j}^\intercal \right| \times \left| r(X_i) - r(x) - \beta^{\star\intercal} (X_i - x) \right| \le k L_2 \overline{\tau}_k^3.$$

We finally obtain that for any $j = 1, \ldots, d$, it holds, with probability $1 - \delta/(2D)$,

$$\left| \mathbb{X}_j^\intercal w \right| \le \sqrt{2k\sigma^2 \overline{\tau}_k^2 \log\left(\frac{16}{\delta}\right)} + k L_2 \overline{\tau}_k^3,$$

and from the union bound, we deduce that, with probability $1 - \delta/2$,

$$\max_{j=1,\ldots,d} \left| \mathbb{X}_j^\intercal w \right| \le \overline{\tau}_k \left(\sqrt{2k\sigma^2 \log\left(\frac{16d}{\delta}\right)} + k L_2 \overline{\tau}_k^2 \right).$$

Proof of eq. (4.16). We show that eq. (4.16) holds with probability $1 - \delta/2$. Define

$$\begin{aligned} \hat{\sigma}_k &= \sum_{i \in \hat{i}_k(x)} (X_i - x)(X_i - x)^\intercal, \\ \hat{\mu}(\tau) &= \sum_{i \in \hat{i}_\tau(x)} (X_i - x). \end{aligned}$$

First, note that

$$\mathbb{X}^\intercal \mathbb{X} = \sum_{i \in \hat{i}_k(x)} (X_i - x)(X_i - x)^\intercal - \frac{1}{k} \hat{\mu}(\hat{\tau}_k) \hat{\mu}(\hat{\tau}_k)^\intercal.$$

Then, using Proposition 4.7, because $k \ge 4\log(4n/\delta)$, with probability $1 - \delta/4$, $\hat{\tau}_k(x) \ge \underline{\tau}_k$, implying that

$$\begin{aligned} \mathbb{X}^\intercal \mathbb{X} &\ge \hat{\sigma}_k - \frac{1}{k} \hat{\mu}(\hat{\tau}_k) \hat{\mu}(\hat{\tau}_k)^\intercal \\ &= \mathbb{E}\left[\hat{\sigma}_k\right] + \left(\hat{\sigma}_k - \mathbb{E}\left[\hat{\sigma}_k\right]\right) - \frac{1}{k} \hat{\mu}(\hat{\tau}_k) \hat{\mu}(\hat{\tau}_k)^\intercal. \end{aligned}$$

Let $u \in \mathbb{R}^d$. We have that

$$\begin{aligned}\left|u^\intercal \hat{\mu}(\hat{\tau}_k)\right|^2 &\leq \|u\|_1^2 \max_{j=1,\dots,d} \left|(\hat{\mu}(\hat{\tau}_k))_j\right|^2 \\ &\leq |S_x| \times \|u\|_2^2 \max_{j=1,\dots,d} \left|(\hat{\mu}(\hat{\tau}_k))_j\right|^2 .\end{aligned}$$

Similarly, we have:

$$\begin{aligned}\left|u^\intercal \left(\hat{\sigma}_k - \mathbb{E}\left[\hat{\sigma}_k\right]\right) u\right| &\leq \|u\|_1^2 \times \left\|\hat{\sigma}_k - \mathbb{E}\left[\hat{\sigma}_k\right]\right\|_\infty \\ &\leq |S_x| \times \|u\|_2^2 \left\|\hat{\sigma}_k - \mathbb{E}\left[\hat{\sigma}_k\right]\right\|_\infty .\end{aligned}$$

Using the variable change $y = x + 2\underline{\tau}_k v$ and that $\underline{\tau}_k \leq \tau_0$, we have that

$$\begin{aligned}\mathbb{E}\left[\hat{\sigma}_k\right] &= n\mathbb{E}\left[(X_1 - x)(X_1 - x)^\intercal \mathbb{1}_{\mathcal{B}(x,\underline{\tau}_k)}(X_1)\right] \\ &= n \int (y - x)(y - x)^\intercal \mathbb{1}_{y \in \mathcal{B}(x,\underline{\tau}_k)} f(y)\, \mathrm{d}y \\ &\geq n b_f \int (y - x)(y - x)^\intercal \mathbb{1}_{y \in \mathcal{B}(x,\underline{\tau}_k)}\, \mathrm{d}y \\ &= n(2\underline{\tau}_k)^{2+d} b_f \int_{v \in \mathcal{B}(0,1/2)} vv^\intercal\, \mathrm{d}v \\ &= n(2\underline{\tau}_k)^{2+d} b_f \left(\int_{-1/2}^{1/2} v_1^2\, \mathrm{d}v_1\right) I_d \\ &= \frac{b_f}{12} n(2\underline{\tau}_k)^{2+d} I_d \\ &= \frac{b_f}{6U_f} \underline{\tau}_k^2 k I_d \\ &\geq \frac{\underline{\tau}_k^2 k}{12} I_d,\end{aligned}$$

using that $U_f/b_f \leq 2$. Consequently,

$$\frac{\|\mathbb{X}u\|_2^2}{\|u\|_2^2} \geq \frac{\underline{\tau}_k^2 k}{12} - |S_x| \left(\left\|\hat{\sigma}_k - \mathbb{E}\left[\hat{\sigma}_k\right]\right\|_\infty + \frac{1}{k} \max_{j=1,\dots,d} \left|(\hat{\mu}(\hat{\tau}_k))_j\right|^2\right).$$

Proposition 4.10 can be applied because

$$\begin{aligned}24nU_f(2\underline{\tau}_k)^d &= 12k \\ &\geq \log\left(\frac{16d^2}{\delta}\right),\end{aligned}$$

which is satisfied whenever C_1 is large. Combined with Proposition 4.11 (our conditions ensure that eq. (4.14) is satisfied), we obtain that, with

probability $1-\delta/4$,

$$\begin{aligned}\frac{\|\mathbb{X}u\|_2^2}{\|u\|_2^2} &\geq \frac{\underline{\tau}_k^2 k}{12} - |\mathcal{S}_x|\left(4\underline{\tau}_k^2\sqrt{\frac{k}{3}\log\left(\frac{16d^2}{\delta}\right)}\right.\\ &\qquad \left. + 2C^2\left(\overline{\tau}_k^2\log\left(\frac{8nd}{\delta}\right) + \frac{L^2 k\overline{\tau}_k^4}{b_f^2}\right)\right)\\ &\geq \frac{\overline{\tau}_k^2 k}{24\times 8}\left(2 - |\mathcal{S}_x| C_3\left(\sqrt{\frac{1}{k}\log\left(\frac{nd}{\delta}\right)} + \frac{1}{k}\log\left(\frac{nd}{\delta}\right) + \frac{\overline{\tau}_k^2 L^2}{b_f^2}\right)\right),\end{aligned}$$

where $C>0$ is a universal constant. To obtain the last inequality we use $\overline{\tau}_k = C_f^{1/d}\underline{\tau}_k$ with $C_f \leq 8$, we choose $C_3 > 0$ large enough and $C_2 > 0$ small enough. Choose C_1 large enough to get that

$$\begin{aligned}C_3\,|\mathcal{S}_x|\,\sqrt{\frac{\log(nd/\delta)}{k}} &\leq \frac{1}{3},\\ C_3\,|\mathcal{S}_x|\,\frac{\log(nd/\delta)}{k} &\leq \frac{1}{3}.\end{aligned}$$

Finally, we obtain the desired result by noting that

$$C_3\,|\mathcal{S}_x|\,\frac{\overline{\tau}_k^2 L^2}{b_f^2} \leq \frac{1}{3}.$$

4.7.5 Proof of Theorem 4.2

We rely on the bias-variance decomposition expressed in eq. (4.9). On the first hand, we have

$$\begin{aligned}|r_k(x) - r(x)| &= \left|\frac{\sum_{i\in\hat{i}_k(x)}(r(X_i)-r(x))}{\sum_{i=1}^n \mathbb{1}_{\mathcal{B}(x,\hat{\tau}_k(x))}(X_i)}\right|\\ &\leq \sup_{y\in\mathcal{B}(x,\hat{\tau}_k(x))}|r(y)-r(x)|\\ &\leq L_1\hat{\tau}_k(x).\end{aligned}$$

Applying Proposition 4.6 we obtain that, with probability $1-\delta/2$,

$$|r_k(x) - r(x)| \leq L_1\overline{\tau}_k.$$

On the other hand, we apply Proposition 4.8 to get that, with probability $1-\delta/2$,

$$|\hat{r}_k(x) - r_k(x)| \leq \sqrt{\frac{2\sigma^2\log(4/\delta)}{k}}.$$

Choose C_1 large enough to get that

$$C\,|\mathcal{S}_x|\,\sqrt{\log\left(\frac{2d}{\delta}\right)} \leq \frac{\sqrt{k}}{3}$$
$$C\,|\mathcal{S}_x|\,d\log\left(\frac{2nd}{\delta}\right) \leq \frac{k}{3}$$

Finally, we obtain the result of interest after noting that

$$C\,|\mathcal{S}_x|\,\overline{\tau}_k^2 L^2 \leq \frac{b_f^2}{3}.$$

Survival Analysis for Securitization | 5

5.1 Introduction

We have seen in §1.1 that nowadays, one of the most important metric a bank has to monitor is the RWA as it directly relates to the capital required and therefore the amount of "dead" assets that do not bring any revenues. While the ability to more accurately estimate p, instead of relying on the rough and very conservative standard models, already brings a regulatory reduction of the RWA by virtue of being a more accurate and lower estimate of the probability of default, it is not sufficient.

Fortunately, it is possible to significantly lower the RWA of the bank even further through the use of active management, which is the primary role of the Portfolio Management team of BNP Paribas. Remember that the RWA is a measure of the proportion of risky assets in the portfolio of BNP Paribas, as such a simple way of lowering the RWA is simply not to have those assets in the portfolio of the bany anymore. The vast majority of financial products exist as a way to transfer risk[128] to other parties who may, because of their risk profile, be indifferent to it and wishing to in exchange be rewarded financially for it. It is therefore desirable in order to reduce the risk of the bank to transfer that risk to somebody else which morally entails selling the credit, i.e. a random value with a variable payoff, to a third party in exchange of a *fair* deterministic price. We will not discuss here what a *fair* price is but for the sake of illustration it is possible to imagine it as the expected value of the payoff under some measure. Of course a single credit is not very appealing as it is, by definition, risky and it is difficult to find a buyer for it, at least for an acceptable price. But BNP Paribas has *many* such credits in its portfolio and is therefore in the position to sell those as a packaged product with lower risk. Such products, usually called asset based security (ABS),[129] are constructed from a pool of credit and, under the assumption that the assets that compose it are not all correlated, will have a lower risk and therefore lower price. Simply put, from the stock of credits c_i in BNP Paribas' portfolio, a product of the form

$$\sum_{i=1}^{n} \omega_i c_i,$$

128: The earliest recorded *financial derivative* is due to Thales of Miletus in 600 BCE, of Thales's theorem fame, who sold a future contract on olives. Thus effectively making Thales the first known derivative trader.

129: A security roughly refer to any tradable and *fungible* financial instrument.

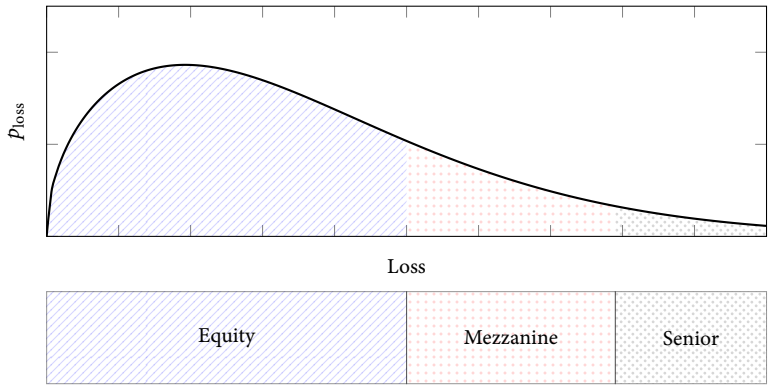


FIGURE 5.1: Tranche structure of an ABS

where ω_i is a set of weights between 0 and 1, is sold to investors, effectively eliminating the RWA corresponding to the portion of the assets sold. In practice, those securitized portfolios sold as ABS involve additional complexities both in the type of assets allowed and the payout structure in order to further satisfy the varied levels of risk appetite of investors. From the simple product presented earlier, more complex derivatives are constructed with different levels of risk and remuneration through an operation known as *tranching*. When the product start experiencing defaults, and therefore incurring losses, instead of distributing the loss uniformly amongst investors, the loss is instead applied by the order of *seniority*, with the *equity* tranche incurring the losses first until it is completely wiped out, and repeating the process for the later tranches as illustrated in fig. 5.1. While it would in theory be possible to create specific portfolios meeting the required risk profile of each investor, it is preferable for purely logistical reasons to create a single portfolio which is then used as a basis for more complex derivatives.[130]

Given the strategic importance for a bank of securitization of credit portfolios, i.e. mathematical objects that entirely rely on the estimation of some probability of default for their pricing, it is crucial to be able to correctly model the default events to guide the construction and pricing of these portfolios.

130: A *derivative* in the financial world refers to a product constructed as a function $P(E(t))$ of some underlying asset $E(t)$.

About this Section

The rest of this chapter is in large part composed of examples inspired by applications at BNP Paribas but not on BNP Paribas data for reasons of ease of access to data, confidentiality as well as reproducibility.

5.2 Portfolio Optimization by Simulation

In chapter 3 we have introduced a flexible estimator of the survival distribution and insisted on the usefulness of a generative model without much more elaboration on the possible applications. We will show here one such application. One significant advantage of our method is the ability to efficiently generate samples of Y, the duration of interest; enabling the possibility to estimate higher order statistics that may depend on Y through a non-trivial process. We present here a toy example designed to mimic the process of securitization in order to motivate this characteristic.

We consider a synthetic dataset of financial entities representing a credit portfolio. For each client i and covariate X_i it is possible to buy an insurance, potentially on a fraction ω_i, of duration d_i which is the maximum protection time of the insurance, for a total price c_i. If the client defaults during the contract duration that is $Y_i \leq d_i$, then the default with loss l_i is entirely covered. We can therefore define the portfolio loss as

$$L(\omega) = \sum_{i=1}^{K} \omega_i (l_i \mathbb{1}_{Y_i \leq d_i} - c_i). \tag{5.1}$$

From the perspective of the client, an optimal portfolio is one that minimizes the potential losses or equivalently maximizes the potential gains. Of course we could minimize $L(\omega)$ in expectation directly but not only is the expectation exactly 0 as this is how the prices c_i are constructed here, but the expected value is often an imperfect metric from the point of view of investors as it does not incorporate any notion of risk.[131] We therefore want to minimize some metric of the risk incurred, of which many variants exist. We chose here to optimize the *expected shortfall*, defined in terms of expected quantiles by

$$\mathrm{ES}_\alpha(\omega) = \int_{-\infty}^{\alpha} F^{-1}_{L(\omega)}(\gamma)\, \mathrm{d}\gamma, \tag{5.2}$$

as it relates to the often-used value-at-risk (VAR), i.e. quantile of level α, and possesses several qualities of interest. More precisely, unlike the VAR, the expected shortfall is a convex and coherent risk measure, that is a monotonous, sub-additive, homogeneous and translationally invariant function. These characteristics not only make the expected shortfall a proper risk measure the of econometrics and microeconomics sense (Artzner et al. [1999]) but more importantly in our specific case a function that can be optimized exactly. This measure is equivalent in the continuous case to the tail conditional expectation and can be seen as *minimizing the expected extremal losses* as seen in fig. 5.2. This objective is also desirable in many other fields such as predictive maintenance or industrial reliability,

131: From a pure expectation point of view, extremely volatile assets are equivalent to non-risky assets, which intuitively *feels wrong*. It is possible to derive a more grounded approach by defining the notion of *utility* (see Markowitz [1952b,a]).

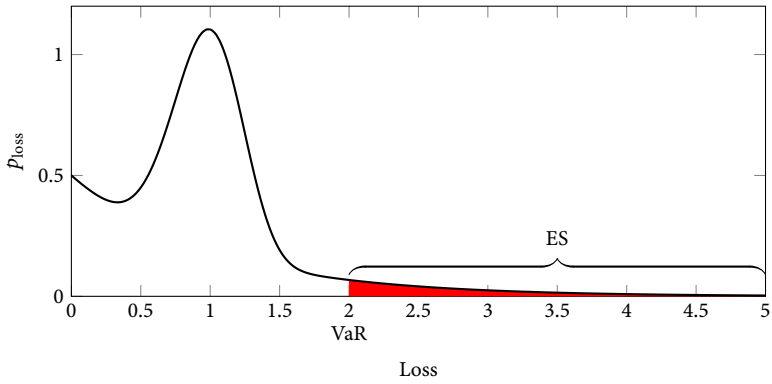


FIGURE 5.2: Expected Shortfall and Value at Risk of a loss distribution.

where minimizing the extreme defects is of particular interest. Not only the previous quantity can be estimated by Monte Carlo simulations using the normalizing flow learned by our method, but it is also possible to directly minimize the previous quantity by solving the convex optimization program

$$\operatorname*{argmin}_{\omega} \int_{-\infty}^{\alpha} F^{-1}_{L(\omega)}(\gamma)\,\mathrm{d}\gamma, \tag{5.3}$$

which can be rewritten as a linear optimization program[132] (see Rockafellar and Uryasev [2002]), if we add constraints on the size of the portfolio P as well as the size of the positions:

$$\begin{aligned} \operatorname*{argmin}_{\omega,\beta} \quad & \beta + \frac{1}{1-\alpha}\int \left[L(\omega)-\beta\right]^{+} p_L(y)\,\mathrm{d}y \\ \text{s.t.} \quad & 0 \leq \omega \leq 1 \\ & \omega^{\intercal} c = P. \end{aligned} \tag{5.4}$$

132: The operation $[\cdot]^+$ can be rewritten in term of linear constraints and the problem stays linear overall.

For simplicity we set here all durations to the same value $d_i = d$ as well as losses $l_i = 1$, and generate default events according to a simplified version of the law introduced here in §3.6.1: we sample 10 base feature vector, after which the 10 resulting vectors are perturbed to form 200 feature vectors, that is

$$\begin{aligned} \tilde{X}_k &\sim \mathcal{U}\left([0,1]^{10}\right) \\ \varepsilon_i &\sim \mathcal{N}\left(0, \mathbb{I}_{10}\right) \\ X_i &= \tilde{X}_{k \bmod 10} + 0.1\varepsilon_i. \end{aligned}$$

The times to defaults and censoring variables themselves still follow the model given previously in §3.6. This simplified model represents the

usually assumed *classes* of risk, such as industries, countries or intrinsic rating, which are often assumed to be similar in terms of default. Prices are chosen as fair prices i.e. as the prices necessary to mitigate the expected losses $c_i = \mathbb{E}[\mathbb{1}_{Y_i \leq d_i}]$. For the pricing we use separate Weibull models learned independently for each $k \in [1, \ldots, 10]$, a common practice in credit rating. We then minimize eq. (5.4) using the reference Weibull models and our model as estimators of the distribution of defaults. The real value of the minimum obtained by both methods is then measured using the true unknown distribution to compute the true realized losses of the portfolio, as we know in this case what the true distribution is which is, of course, not the case in practice.

The true expected shortfall of the standard optimal portfolio is 8.1452 while the optimal portfolio formed using the survival flow estimates achieves an expected shortfall of −0.314, which translates into an economic *gain* in this scenario. Purely on this metric, the more granular and accurate samples from the normalizing flow model results in a significantly better value of the objective. While the expected shortfall is a good objective with interesting mathematical characteristics as well as a real economic interpretation, investors are in the end interested in the real possible returns. As seen in fig. 5.3, the potential losses from the optimal portfolio obtained by means of minimizing the expected shortfall derived from the reference distribution, i.e. the distribution used for the pricing, are significantly higher than those obtained using samples from a normalizing flow based estimate of the survival distribution. In this toy example, we clearly observe better losses (or gains depending on the point of view) when sampling from the survival flow distribution.

While a purely synthetic toy problem, this serves as an example of how a better estimator of the survival distribution than those generally in use in banking institutions can help construct significantly better portfolios that should also perform better during periods of intense stress by abandoning the potentially damaging hypothesis of independence between companies.

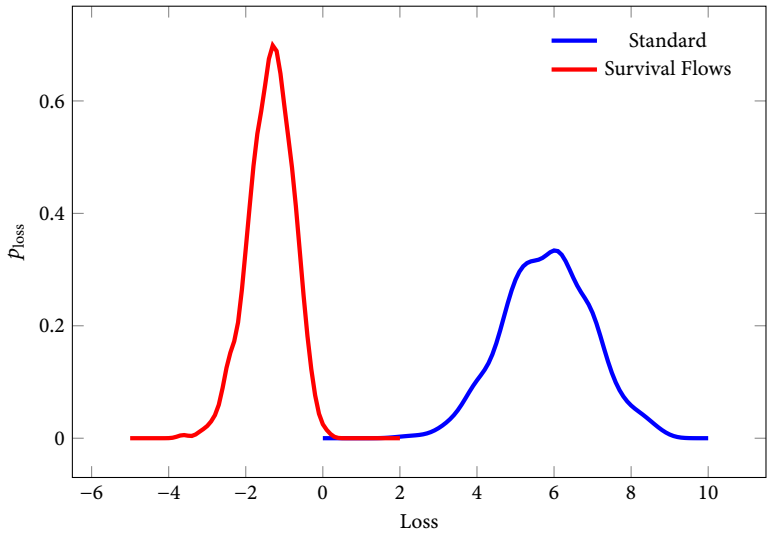


FIGURE 5.3: Realized losses for the optimal portfolios using the reference and normalizing flows estimates.

5.3 Deep Bayesian Survival Analysis

The previous example, while a good motivation for better estimators of the survival, is however mostly a worthwhile exercise from the point of view of the buyer and not the originator of the securitization portfolio, BNP Paribas.

As eluded earlier, BNP Paribas is able to leverage its extensive portfolio to construct products satisfying the needs of their clients, which taking the optimization view of eq. (5.4), entails adding constraints to match the risk profile of the potential clients. One usual limitation that, however, greatly hinders the potential RWA savings of BNP Paribas by limiting its ability to sell part of its portfolio is the limitation of permissible ratings. While the very act of securitization is meant to limit risk of the overall portfolio, most clients still demand that the portfolio only contains highly rated and low-risk assets, making it impossible to sell the more average and moderately risky assets. Given the limited pool of highly rated companies and the required size of the overall portfolio, the only solution in order to satisfy all client and regulator constraints is to incorporate a very significant fraction of the few but large lines of credits from highly rated[133] clients. One side effect of these constraints being that fewer large clients are included instead of a multitude of smaller ones, thus effectively reducing the diversification of the resulting portfolio and increasing the overall risk. Worse, given the size of the credit lines involved, the risk is increased in the tail of the distribution of the losses and therefore represents an unacceptable stress scenario, all while paradoxically bringing very little profits to the

133: Often sovereign.

buyers as the prices are calculated from the ratings that compose the portfolio. In order to convince the potential clients to review their policy concerning middle-of-the-road ratings, it is possible to build a model that incorporates the uncertainty on the ratings and show that given this uncertainty it is recommended to diversify the allowed ratings in order to reduce the heaviness of the tail of the distribution.

We first start by modelling the distribution of the true probabilities of defaults given the ratings in order to show that *good* ratings are highly uncertain in their estimation, given the very few observations of defaults. A good way of incorporating uncertainty in the estimation of the probabilities of defaults for each rating is simply to treat those as random variables, in a Bayesian setting. This also enables us to introduce additional knowledge in the priors and model in order to alleviate the problem of few observations. Instead of determining the probability of defaults of each rating by simply taking

$$p_i = \frac{d_i}{n_i},$$

where d_i is the number of ratings i which defaulted and n_i the total number of assets of ratings i in the portfolio, we can instead treat each p_i as a random variable such that,

$$\begin{aligned} &\alpha_i \in \mathbb{R}_+,\ \beta_i \in \mathbb{R}_+ \\ &p_i \sim \text{Beta}(\alpha_i, \beta_i) \\ &d_i \sim \text{Binomial}(n_i, p_i). \end{aligned}$$

However, as the good ratings generally never see any defaults, and therefore more often than not are all equals to a probability of 0 most years, it is not rare that by pure chance no defaults happen for the ratings 3 and 4 but a single one does for the rating 2, for example. In those cases a naive estimation would result in believing that the probability of default is higher for the ratings 2 than 3 and 4. If we, however, make the assumption that the analysts in charge of the ratings are actually correct, which seems more likely than the ratings being the complete opposite of reality, we can instead force into the model the assumption that the ratings are strictly increasing by modelling them as such, e.g.:

$$\begin{aligned} &\alpha \in \mathbb{R}_+^K \\ &\Delta p \sim \text{Dirichlet}(\alpha) \\ &d_i \sim \text{Binomial}\left(n_i, \sum_{j=1}^{i} \Delta p_j\right). \end{aligned}$$

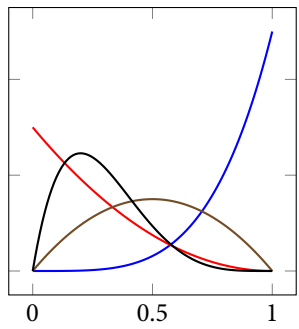


FIGURE 5.4: Several Beta distributions.

The previous example was applied on BNP Paribas' internal data in order to show that the uncertainty in the estimation of the probability of the

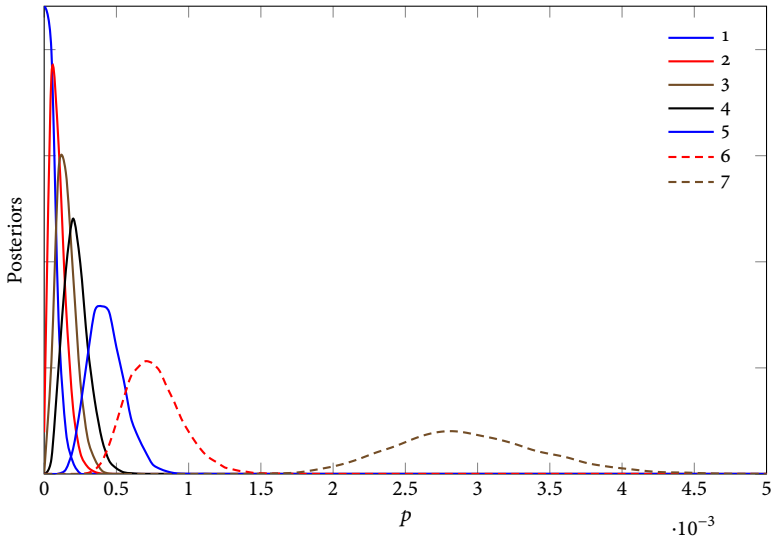


FIGURE 5.5: Posterior distribution of the probabilities of defaults.

good ratings was too high to be useful for the derivation of strict guidelines. In fig. 5.5 we give the results for the first 7 aggregated ratings, that is we aggregate R-, R and R+ into a single rating R, and we clearly see the posterior distributions of the possible probabilities of default overlap nearly entirely for the first few ratings. Given how few observations we have available for estimation, and therefore how spread the posterior distributions are, it is dubious to arbitrarily decide that ratings 1–3 are good and 4 is not. From these posterior distributions, it is then possible to proceed as in §5.2 in order to derive an optimal portfolio which takes into account the uncertainty around the estimated probabilities of defaults by solving the problem with the business constraints incorporated, i.e.

$$\begin{aligned} \underset{\beta,\omega}{\text{argmin}} \quad & \beta + \frac{1}{n(1-\alpha)} \sum_{i}^{n} [\omega_i (s_i e_i - c_i) - \alpha]^+ \\ & \omega^\intercal c = P \\ & \omega_i c_i \geq 30 s_i \\ & \omega \leq \delta \\ & s \in \{0,1\}^K \end{aligned} \tag{5.5}$$

where the constraint $\omega_i c_i \geq 30 s_i$ encodes the fact that we only want to add an element to the portfolio if the total contribution is large enough to justify the fixed cost of doing so. Because of the introduction of the binary variables s_i, the problem is not linear anymore but instead a mixed integer linear programming (MILP) problem which, while harder to solve,

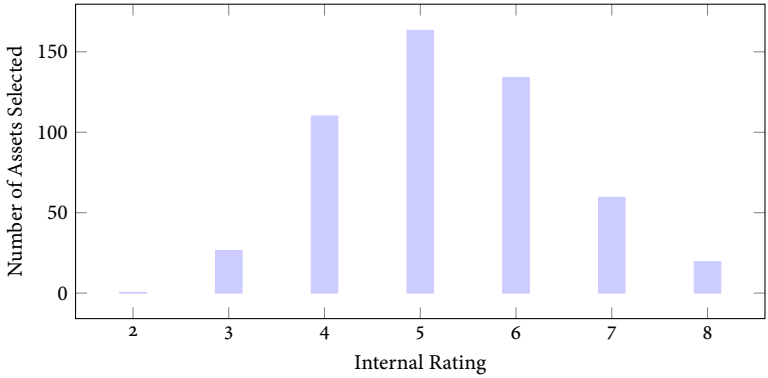


FIGURE 5.6: Assets selected by rating.

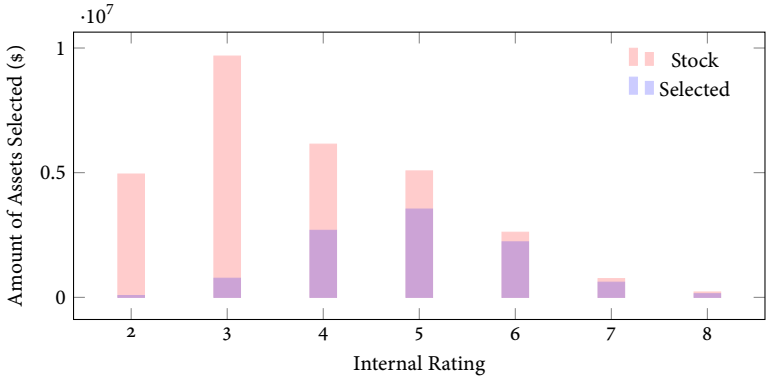


FIGURE 5.7: Portfolio Selected compared to existing portfolio.

is still tractable. Unsurprisingly, as it is what we set out to show, we see in fig. 5.6 that the composition of the portfolio is heavily weighted toward the more average ratings as their true probability of default is better estimated and their number and diversity makes it easier to assemble a low-risk portfolio. More importantly from the point of view of BNP Paribas, we see in fig. 5.7 that this results in most of the existing stock of assets rated worse than 3 being securitized and sold, with the reduction in RWA it entails. While interesting, the previous approache still relies on transforming the problem of estimating defaults into a binary classification problem, which is exactly what we set out to *not* do by reframing it as a time-to-event prediction problem in the rest of this manuscript. The Bayesian modelling approach presented earlier is appealing because it gives us the ability to easily obtain uncertainty estimates as well as make efficient use of the data by incorporating as much knowledge as possible through a careful choice of priors as well as by hierarchically pooling the parameters. This approach is still possible for survival regression: in chapter 3 we introduced survival normalizing flows, a generative model of the time-to-event with a tractable

likelihood and, incidentally, the ability to sample as well as to compute the likelihood are the only two requirements for Markov chain Monte Carlo (MCMC) estimation of Bayesian models to be possible.

In the credit setting, it is usual to develop different but related models for each industry category or jurisdiction in order to predict the probability of default, as two companies with otherwise identical characteristics X may have wildly different default behaviours due to idiosyncrasies in their sector or country of residence. The naive approach to the estimation of K different distributions of time-to-events of these K specific subpopulations results in suboptimal results as the effective sample size available for training is drastically reduced. We can, however, make the hypothesis that inside those subpopulations the different distributions of time-to-events are fairly similar and share most of their characteristics and parameters. While it would be possible, as is often done in deep learning, to treat the problem as a multi-task learning problem with hard-sharing of the first few layers of the neural networks parametrizing the flows, we can instead use the well-principled approach of *multilevel hierarchical modelling* (Gelman [2006]; Raudenbush [1988]). In chapter 3 we modelled the time to event Y of the population i as following some distribution parametrized by a normalizing flow itself parametrized by a fixed θ_i,

$$Y \mid \text{pop } i \sim \text{NF}(\theta_i).$$

Instead, we can treat θ_i as some unknown observation from a prior distribution, such that all the θ_i share information through their common latent distribution, that is

$$\begin{aligned} \theta_i &\sim \mathcal{N}(\mu, \Sigma) \\ Y \mid \text{pop } i &\sim \text{NF}(\theta_i). \end{aligned}$$

The same is, of course, also possible when conditioned on X

$$\begin{aligned} \theta_i &\sim \mathcal{N}(\mu, \Sigma) \\ Y \mid X, \text{pop } i &\sim \text{NF}(\theta_i, X). \end{aligned}$$

This hierarchical approach is particularly interesting when each population taken independently has too few observations for an accurate estimator to be trained. This is for example the case when expanding into new markets: mostly similar data is already available for the markets where the bank already has an historical presence but only a few observations of the new market exist for the first few months or years. On one hand, there are not enough points to train a new model while, on the other hand, reusing models trained on different populations is not recommended as the distributions do not have any reason to be exactly the same. The

assumption that the distributions are *similar* is, however, fairly sensible and it makes sense to reuse past observations as prior knowledge for our new model.

5.4 Conclusion

The ability to estimate the conditional probability of default or, even better, the distribution of the time-to-defaults is crucial to the correct operation of a bank such as BNP Paribas. The probability of default is used to define internal ratings, to decide whether to accept or not loan applications, to price products ranging from simple insurances to complex derivatives or to estimate counterparty risk and much more. In this last chapter, we have chosen not to give any example of these problems of estimation of a time-to-event in the specific setting of finance as those problems, except for the name of the event, match exactly the examples one can find in the medical or industrial world and described at length in chapters 2 to 4. *Securitization*, however, is a very specific financial application and the core competency of the Portfolio Management team at BNP Paribas. Given the ever-increasing regulatory requirements of each new BASEL framework, the need to reduce RWA has increased significantly and, in many ways, is now a more important metric to the bank that the economic gain itself. In order to survive, banks have to drastically change their operating model and improve their capability to estimate credit risk and perform stress tests. Regulators are not satisfied with simply meeting capital requirements anymore and increasingly demand proofs through stress tests of the resilience of banks and other financial actors to adverse financial conditions. As such, generative models which were once seen as an expensive bonus, are now crucial tools necessary to the day-to-day operation of a bank. We have presented in this chaper several applications of generative models, either alone or part of a more complex Bayesian framework, and shown how these models can be exploited to improve every step of the decision process of a portfolio management and securitization team. Thanks to new performant tools for probabilistic programming (Salvatier, Wiecki, and Fonnesbeck [2016]; Ge, Xu, and Ghahramani [2018]; Cusumano-Towner et al. [2019]) and abundant compute power, generative methods are now in reach of most practitioners. Adapting recent advances such as diffusion models (Nichol and Dhariwal [2021]) to the survival setting therefore seems a worthwhile effort.

Conclusion and Perspectives | 6

Predicting the time until an event happens, or *time-to-event*, is a fundamental problem in many fields such as medicine, reliability theory, and finance. Despite the apparent simplicity of the task: regressing a positive random variable Y; the subject quickly proves to be complicated by the presence of censoring i.e. the inability to completely observe the variable Y and instead only observe $\min(Y, C)$ and $\mathbb{1}_{Y \leq C}$, a partial observation and a censoring indicator. Given the wide range of important, and in many case life saving, applications; *survival analysis* has given rise to an impressive and wide-ranging body of research, spearheaded by the statistical community.

In this thesis we presented a different theoretical approach to survival analysis through the scope of machine learning and more precisely the ERM approach. We showed how, through a relatively simple in practice reweighing of the observations, it is possible to adapt the ERM objective in order to deal with the presence of censoring. We then derived nonasymptotic and nonparametric upper bounds on the excess risk that match the bounds one can obtain without censoring. Not only do those theoretical results justify reusing many powerful regression methods from the machine learning literature, but we also showed empirically that this approach manages to match or even beat the state-of-the-art in survival analysis when the task of interest is purely predictive; which is increasingly the case with the democratization of machine learning in the industry. Some interrogations on the method, however, do remain: as our loss is now estimated and depends on the choice of some hyperparameters itself, it becomes necessary to select the best possible hyperparameters for the estimation of the loss. However, traditional cross-validation (CV) approaches cannot be directly applied as we do not have access to the real loss on which to measure our performance. It may therefore be necessary to study in future work an approach based on marginalizing those parameters (Brault et al. [2019]). Additionally, our results only encompass the static setting where the covariates X are supposed to be independent of time which is, both in general and more specifically the survival setting, a very limiting constraint. Extending the framework introduced in this thesis to the dynamic setting where the covariates $X(t)$ are allowed to vary with time is therefore

a natural and import direction for future work.

While the experiments on the IPCW ERM framework showed that even simple estimators of the survival are enough for the reweighing to yield good results, better weights still lead to better results. We therefore introduced a highly flexible neural based estimator of the survival based on continuous normalizing flows and showed that this new estimator outperformed or matched the current state-of-the-art on censored regression and ranking tasks. Our estimator has the particularity, compared to most of the competition, to be a generative model with a flexible likelihood which enables novel applications beyond scoring and estimating the survival function, such as the efficient generation of new samples or its use in Bayesian estimation. This flexibility, however, comes at the cost of a greatly increased computational overhead which should be mitigated by further research on efficient neural architectures for survival analysis, or even the pursuit of different approaches adapted from the generative diffusion process literature.

This computational cost can be partially solved by reducing the dimension of the covariates beforehand, a task known as *variable selection*. In order to solve this problem, we showed how to estimate the gradient when it is supposed sparse by formulating the gradient learning problem as a LASSO local linear regression problem with k-NN averaging. We gave nonasymptotic upper bounds on the error of the resulting sparse estimate of the gradient as well as the estimate of the regression function itself. This k-NN and LASSO approach offers a robust and easy to calibrate procedure for the estimation of the gradient from which we can deduce the most important variables, either globally by forming the expected gradient outer product as is done in the single or multi-index approach, or locally by incorporating the gradient information inside the splitting procedure of a random forest.

Finally, we made use of the methods introduced in this thesis on the financial task of *securitization*, which involves the creation of a portfolio of loans, and showed that the machine learning approaches proposed lead to significantly improved results for BNP Paribas. Survival normalizing flows are used for their generative property to sample time-to-defaults in order to solve an empirical portfolio optimization problem whose solution greatly outperforms the standard approach. Additionally, we made use of both the generative and tractable likelihood aspects of normalizing flows to motivate their inclusion inside a Bayesian graphical model, playing the role of a highly tractable black box distribution.

Some Useful Bounds | A

Most of the results and proofs presented in the earlier chapters involve in some way or another tail bounds of distributions or empirical processes. There are too many results to list, and as this is not the goal of this thesis we will not give any proofs. We will, however, quickly restate some standard bounds so that the reader does't have to reach for another reference book in order to follow some of the proofs. We heartily recommend Wainwright (2019) to the readers who want to have a very exhaustive overview of the bounds one can derive in the non-asymptotic and non-parametric setting.

Proposition A.1 (Chernoff's bounding technique.)**.** *Recall Markov's inequality, for any L^1 integrable random variable X we have*

$$\mathbb{E}[X \geq t] \leq \frac{\mathbb{E}[X]}{t}, \ \forall t > 0. \tag{A.1}$$

In particular, if X admits a moment generating function on a neigbourhood $[0, b]$ then

$$\log \mathbb{P}\left[(X - \mu) \geq t\right] \leq \inf_{[0,b]} \left(\log \mathbb{E}\left[e^{\lambda(X-\mu)}\right] - \lambda t\right). \tag{A.2}$$

The technique of introducing an auxialiary variable to be optimized is commonly refered to as Chernoff's bounding technique *and yields many inequalities as a special case.*

Definition A.1. A random variable X with mean $\mu = \mathbb{E}[X]$ is said to be σ sub-Gaussian if there exists $\sigma > 0$ such that

$$\mathbb{E}\left(e^{\lambda(X-\mu)}\right) \leq e^{\sigma^2\lambda^2/2}$$

Proposition A.2 (Hoeffding's inequality)**.** *Let (X_i) for $i = 1, \ldots, n$ be independent σ_i sub-Gaussians of means $\mu_i < \infty$, then for all $t > 0$*

$$\mathbb{P}\left(\sum_{i=1}^{n}(X_i - \mu_i) \geq t\right) \leq \exp\left(-\frac{t^2}{2\sum_{i=1}^{n}\sigma_i^2}\right).$$

Theorem A.3 (Characterizations of sub-Gaussianity)**.** *For any centered random variable X, all the following points are equivalent:*

1. *(σ sub-Gaussianity) There exists $\sigma > 0$ such that*

$$\mathbb{E}\left[e^{\lambda X}\right] \leq \exp\left(\frac{\lambda^2\sigma^2}{2}\right) \qquad \forall\lambda \geq 0.$$

2. *There exists $c \geq 0$ and $Z \sim \mathcal{N}(0, \sigma^2)$ such that*

$$\mathbb{P}\left(|X| \geq t\right) \leq c\mathbb{P}\left(|Z| \geq t\right) \qquad \forall t \geq 0$$

3. *There exists $c \geq 0$ such that*

$$\mathbb{E}\left[X^{2k}\right] \leq \frac{(2k)!}{2^k k!} c^{2K} \qquad \forall k \in \mathbb{N}^+$$

4. *There exists $\sigma > 0$ such that*

$$\mathbb{E}\left[e^{\frac{\lambda X^2}{2\sigma^2}}\right] \leq \frac{1}{\sqrt{1-\lambda}} \qquad \forall\lambda \in [0, 1[$$

Definition A.2. A random variable X with mean $\mu = \mathbb{E}[X]$ is said to be (ν, α) sub-Exponential if there exists $(\nu, \alpha) \in \mathbb{R}_+ \times \mathbb{R}_+$ such that

$$\mathbb{E}\left[e^{\lambda(X-\mu)}\right] \leq \exp\left(\frac{\nu^2\lambda^2}{2}\right) \qquad |\lambda| < \frac{1}{\alpha}.$$

In particular, all sub-Gaussian variables are also sub-Exponential.

Proposition A.4 (Sub-Exponential tail bound.)**.** *Suppose that X is (ν, α) sub-Exponential, then*

$$\mathbb{P}\left(X - \mu \geq t\right) \leq \begin{cases} \exp\left(-\frac{t^2}{2\nu^2}\right) & \text{for } 0 \leq t \leq \frac{\nu^2}{\alpha} \\ \exp\left(-\frac{t}{2\alpha}\right) & \text{for } t \geq \frac{\nu^2}{\alpha} \end{cases}$$

Proposition A.5 (Bernstein-type bound)**.** *For any random variable X such that $\mu = \mathbb{E}[X]$ and $\sigma = \mathbb{E}\left[X^2\right] - \mu^2$ satisfying the Bernstein's condition:*

$$\left|\mathbb{E}\left[(X-\mu)^k\right]\right| \leq \frac{1}{2}k!\sigma^2 b^{k-2} \qquad \forall k = 2, 3, \cdots$$

we have

$$\mathbb{E}\left[e^{\lambda(X-\mu)}\right] \leq \exp\left(\frac{\lambda^2\sigma^2}{2(1-b|\lambda|)}\right) \qquad \forall|\lambda| \leq \frac{1}{b}$$

as well as

$$\mathbb{P}\left(|X-\mu| \geq t\right) \leq 2\exp\left(-\frac{t^2}{2(\sigma^2+bt)}\right) \qquad \forall t \geq 0.$$

Theorem A.6 (Characterization of sub-Exponential variables.)**.** *For a centred random variable X, all the following are equivalent:*

1. *((ν, α) sub-Exponential) There exists $(\nu, \alpha) \in \mathbb{R}_+ \times \mathbb{R}_+$ such that*

$$\mathbb{E}\left[e^{\lambda(X-\mu)}\right] \leq \exp\left(\frac{\nu^2\lambda^2}{2}\right) \qquad |\lambda| < \frac{1}{\alpha}.$$

2. *There exists $c > 0$ such that*

$$\mathbb{E}\left[e^{\lambda X}\right] < \infty \qquad \forall |\lambda| < c$$

3. *There exists $c_1, c_2 > 0$ such that*

$$\mathbb{P}\left(|X| \geq t\right) \leq c_1 e^{-c_2 t} \qquad \forall t > 0$$

4. *The following constants exist and are bounded:*

$$\sup_{K \geq 2}\left(\frac{\mathbb{E}\left[X^k\right]}{k!}\right)^{1/k} < \infty$$

Definition A.3. A sequence of random variables (Y_i) adapted to a filtration $(\mathcal{T}_i)$ is said to be a martingale if for all $k \geq 1$ we have

$$\mathbb{E}\left[|Y_k|\right] \leq \infty$$
$$\mathbb{E}\left[Y_{k+1} \mid \mathcal{T}_k\right] = \mathbb{E}\left[Y_k\right]$$

Theorem A.7. *Let (Y_k) be a martingale with respect to the filtration $(\mathcal{T}_k)$ and suppose that almost surely*

$$\mathbb{E}\left[e^{\lambda(Y_{k+1}-Y_k)} \mid \mathcal{T}_k\right] \leq e^{\lambda^2\nu_k^2/2} \qquad \forall |\lambda| < \frac{1}{\alpha_k}$$

then we have the following results

1. $\sum_{k=1}^n (Y_{k+1} - Y_k)$ *is* $\left(\sqrt{\sum_{k=1}^n \nu_k^2}, \max_{k=1,\dots,n} \alpha_k\right)$ *sub-Exponential*
2. *The sum of martingale differences satisfies*

$$\mathbb{P}\left(\left|\sum_{k=1}^n (Y_{k+1} - Y_k) \geq t\right|\right) \leq \begin{cases} 2\exp\left(-\frac{t^2}{2\sum_{k=1}^n \nu_k^2}\right) & 0 \leq t \leq C \\ 2\exp\left(-\frac{t}{2\max_{k=1,\dots,n} \alpha_k}\right) & t \geq C \end{cases}$$

where

$$C = \frac{\sum_{k=1}^n \nu_k^2}{\max_{k=1,\dots,n} \alpha_k}$$

Corollary A.8 (Azuma-Hoeffding's inequality). *Let (Y_k) be a martingale with respect to the filtration $(\mathcal{F}_k)$ and suppose that there exists $(a_k, b_k$ such that almost surely*

$$Y_{k+1} - Y + k \in [a_k, b_k] \qquad \forall k \in \mathbb{N}^+.$$

Then for all $t \geq 0$ we have

$$\mathbb{P}\left(\left|\sum_{k=1}^{n}(Y_{k+1} - Y_k) \geq t\right|\right) \leq 2\exp\left(-\frac{2t^2}{\sum_{k=1}^{n}(b_k - a_k)^2}\right)$$

Corollary A.9 (Bounded differences inequality). *Let f a function satisfying the condition*

$$\left|f(x) - f(x^{\leftarrow x'_k})\right| \leq L_k \qquad \forall k = 1, 2, \ldots, n,\ \forall x, x' \in \mathbb{R}^n$$

where $x^{\leftarrow x'_k} = [x_1, \ldots, x_{k-1}, x'_k, x_{k+1}, \ldots, x_n]$, then

$$\mathbb{P}\left(\left|f(X) - \mathbb{E}\left[f(X)\right]\right| \geq t\right) \leq 2\exp\left(-\frac{2t^2}{\sum_{k=1}^{n} L_k^2}\right) \qquad \forall t \geq 0$$

Theorem A.10. *Let $(X_1, \ldots, X_n)$ be a vector of independant and identically distributed standard Gaussian variables, and $f : \mathbb{R}^n \mapsto \mathbb{R}$ be L-Lipschitz. Then $f(X) - \mathbb{E}[f(X)]$ is M sub-Gaussian with $M \leq L$ i.e.*

$$\mathbb{P}\left(\left|f(X) - \mathbb{E}[f(X)]\right| \geq t\right) \leq 2\exp\left(-\frac{t^2}{2L^2}\right) \qquad \forall t \geq 0.$$

Lemma A.11. *Suppose that $f : \mathbb{R}^n \mapsto \mathbb{R}$ is differentiable, then for any convex $\varphi : \mathbb{R} \mapsto \mathbb{R}$:*

$$\mathbb{E}\left[\varphi\left(f(X) - \mathbb{E}\left[f(X)\right]\right)\right] \leq \mathbb{E}\left[\varphi\left(\frac{\pi}{2}\left\langle \nabla f(X), Y\right\rangle\right)\right],$$

where $X, Y \sim \mathcal{N}(0, \mathbb{I}_n)$ independent standard gaussians.

An History of Survival Analysis | B

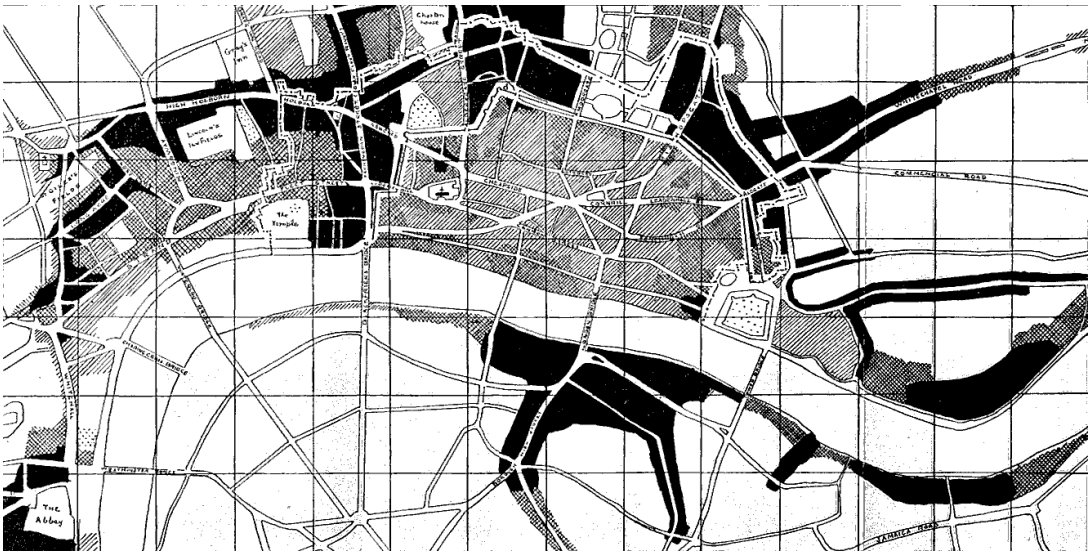


Figure B.1: Geographical Distribution of the Great Plague, London (1606).

B.1 Counting deaths

In the introduction we argued that most of the historical examples of survival analysis treat the problem as an estimation problem whose goal is to *understand* the causes and effects of some phenomenon, usually a disease, on the life expectancy; as opposed to the less inquiring question of prediction. We give here two early examples of survival analysis that both laid the foundations of techniques we ourselves made use of in this thesis: survival analysis as a counting process, and survival analysis as a dynamical process.

B.1.1 The Plague

In 1345, the Mongol army of the Golden Horde laid siege to the Genoese trade port of Kaffa in Crimea. While this event is today largely forgotten,

Age Interval	Proportion of Deaths	Proportion of Survivors Until Start
0–6	0.36	1.00
7–16	0.24	0.64
17–26	0.15	0.40
27–36	0.09	0.25
37–46	0.06	0.16
47–56	0.04	0.10
57–66	0.03	0.06
67–76	0.02	0.03
77–86	0.01	0.01

Table B.1: Of the number of inhabitants.

its consequences had a long-lasting impact on modern civilization. The Mongol army of Jani Beg didn't only bring with it warfare, it also brought a more insidious enemy: the bubonic plague. By fleeing from Kaffa, the Genoese merchant ships brought with them the deadly disease, spreading it along the trade routes of Medieval Europe and causing the death of nearly half of the European population and while no outbreak of the same magnitude ever impacted Europe since then, the plague became a fact of life for most of Europe. In 1665, England experienced its last great outbreak in the Great Plague of London. While the plague of 1345 was mostly characterized by its lack of data, scientific methodology had flourished and the foundations of what would one day become the field of statistics had already been laid. In 1661, in the midst of the last Plague outbreaks, John Graunt, a Londonian haberdasher, was tasked by a group of merchants of London to collect data on the viability of commercial enterprises, or in plain English: to perform market research on the number of potential customers. In 1661, a potential customer was first and foremost a living customer and it would have been foolish to start a new commercial endeavour in a borough without sufficient population. While data was then, to our standards, scarce and often incomplete; one may be surprised to learn that demographic data was meticulously recorded and archived, often by the clergy. By tallying up the records of local parishes, the *City of London* was able to produce accurate records of the demography both in birth and deaths in the form of Bills of Mortality. table B.3 is one such *bill of mortality*, as it would have been available the Graunt during his work. Later, other information such as the supposed *cause of death* or *age of death* would be added. Graunt wasn't a statistician or mathematician by any standards but was able to interpret the data from the bill of mortality from intuition and after making a few hypotheses was able to build the following table B.1 From the table B.1, Graunt was able to derive table B.2, the probability of dying in a given interval provided you had survived until then. Not only, unknown to him or any of his peers, was Graunt able to tabulate $S(x) = 1 - F(x)$ (second column of table B.1), the survival

Age Interval	Probability of Dying If Alive at Start
0–6	0.06
7–16	0.0375
17–26	0.0375
27–36	0.036
37–46	0.0375
47–56	0.04
57–66	0.05
67–76	0.0667
77–86	0.1

Table B.2: Rate of death by age interval.

function of a Londoner, but also the hazard rate for a specific interval. Of course, the notion of cumulative distribution function didn't exist yet, and Newton[137] wouldn't invent calculus until 60 years later, making any insights into the *instantaneous rate* of death impossible. Graunt's results were, however, more than enough to derive insights: while it may seems obvious today, he was amongst the first to treat census data as ordered data, making the assumption that people close in age are more related than people of completely different ages enabling him to derive rate of dying yearly for each age. From these insights Graunt was not only able to estimate the life expectation of a Londoner to be,

137: Of course, as a Frenchman it is my duty to remind the reader that Leibniz's calculus *obviously* came first.

$$\begin{aligned}\mathbb{E}(X) &= 3 \times .36 + 10.5 \times .24 + 20.5 \times .15 \\ &\quad + 30.5 \times .09 + 40.5 \times .06 + 50.5 \times .04 \\ &\quad + 60.5 \times .03 + 70.5 \times .02 + 80.5 \times .01 \\ &= 18 \text{ years}\end{aligned}$$

but more importantly to notice that people were dying at the same rate whether they were 20 or 50 years old: the primary cause of death was therefore not attributable to age and couldn't be natural: it was the plague. John Graunt was able to derive the previous results having only access to aggregated population data which by chance made the task easier and tractable using the tools at disposition at the time. Indeed, by deciding to study age intervals and by recording deaths in each interval, Graun unknowingly recorded *at-risk* individuals and *events* therefore bypassing the problem of censoring that we will introduce later. The formalization of Graunt's insight is exposed in §1.2 and §2.1 through the Kaplan-Meier and Nelson-Aalen estimators, and more generally the product-integral approaches resulting from the counting process view of survival analysis. Graunt's archaic, but impressively enlightened, form of data analysis of mortality data was enough to answer the question being asked. As knowledge increases so do the complexity of the questions we want answered. While Graunt and his merchant sponsors were perfectly content

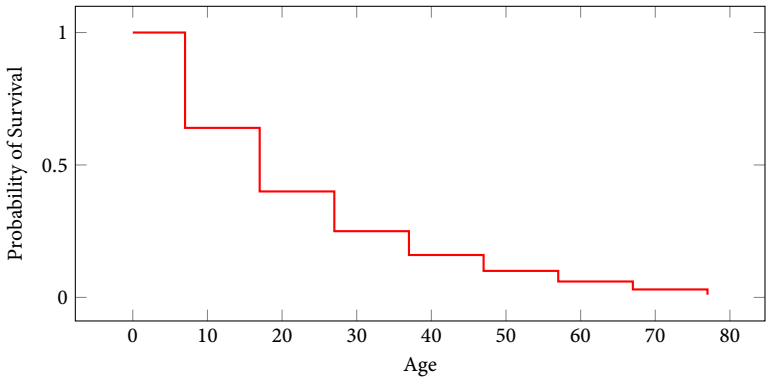


FIGURE B.2: Survival function of a Londoner in 1606.

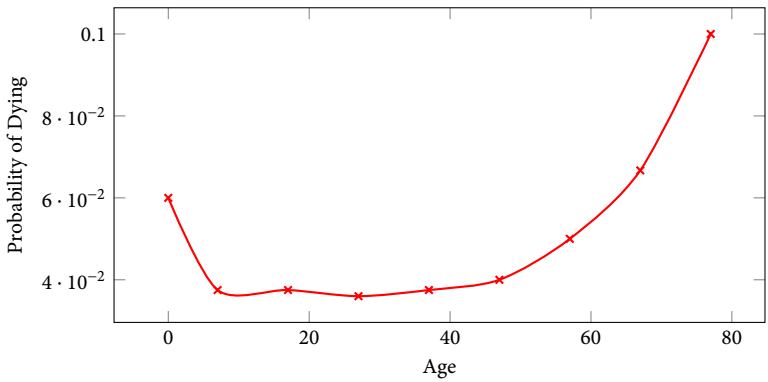


FIGURE B.3: Probability of dying within the year at a given age for a Londoner in 1606.

only knowing the proportion of people alive at each age, surely survival data can be put to better use.

B.1.2 The Smallpox.

The plague may have, in the mind of many, become associated with the idea of deadly pandemics, especially to someone born in recent times, after the advent of vaccination and other prophylactic measures. It is therefore not a surprise that most would not suspect that until very recently diseases that were far more deadly and prevalent were commonplace enough to be part of everyday life. While Graunt and his contemporaries didn't have any means of fighting the plague other than vague yet in a way surprisingly accurate theories around *miasmas*, leading to politics of enhanced sanitation and what we would today call *social distancing*, there didn't exist any medical solution to the disease, making counting deaths the only possible application of survival analysis. This wasn't the case for smallpox, the other great disease of the time, a bigger and more familiar killer. People knew that unlike the black death, smallpox could be fought: indeed if you already suffered from it, you couldn't become ill to it again. This knowledge led the Ming dynasty to encourage *inoculation* or *variolation*: the process of exposing a subject to smallpox (by scraping a smallpox blister before scraping the subjects' skin) in a controlled manner. British and Dutch merchants, learning from this practice reported it back in the continent. Inoculation is not however harmless, it is after all exposition to a live virus with an extremely high case fatality (case fatality, or the proportion of infected people who die, was estimated to around 30% for normal infections), even if through trial and error the advocates of inoculation discovered that proper inoculation had a much lower case fatality (of around 2–3%) due to a lower and controlled viral dose.

While it seemed to many that inoculation did in fact work; considering the fact that nobody could understand why, it is not surprising that the idea of inoculating large swathes of the population was met was great resistance. It was therefore necessary to prove that the risk taken was indeed beneficial and outweighed the potential side effects.

While by the late 1700s, most of Europe had already adopted variolation as a means to fight small pox (with the acceleration of the concentration of population in cities, it was now estimated that 1/10 of the population would eventually succumb to the smallpox.), France seemingly resisted or at least proved to be sceptical to the idea of mass variolation of the population. This fear of variolation was not limited only to the "uneducated" masses and cannot be entirely blamed on obscurantism as virulent opposition often came from intellectuals still revered today such as Voltaire:

Figure B.4: Smallpox Vaccination (1934), Ota Chou

> It is inadvertently affirmed in the Christian countries of Eu-

Weeks.	Days of the Month.	Chriſt.	Bur.	Pla.	Par. infec.	Weeks.	Days of the Month.	Chriſt.	Bur.	Pla.	Par. infec.
1	Dec. 26	100	116	5	5	28	July 3.	109	110	25	12
2	January 2.	117	151	6	5	29	10.	111	134	33	18
3	9.	130	138	4	4	30	17.	115	146	50	22
4	16.	124	138	3	2	31	24.	96	140	46	26
5	23.	143	121	6	4	32	31.	132	178	66	29
6	30.	124	101	3	2	33	Auguſt 7.	131	181	67	29
7	Febr. 6.	122	105	5	5	34	14.	141	197	75	33
8	13.	131	118	7	6	35	21.	133	189	85	28
9	20.	126	109	12	6	36	28.	125	207	85	29
10	27.	102	117	9	8	37	Septem. 4.	123	241	116	32
11	March 6.	110	98	7	4	38	11.	134	216	105	28
12	13.	126	137	9	7	39	18.	121	214	92	36
13	20.	123	133	14	11	40	25.	132	204	87	35
14	27.	134	123	17	8	41	October 2.	121	256	141	40
15	April 3.	123	114	13	9	42	9.	134	218	106	38
16	10.	132	145	27	11	43	16.	142	227	117	37
17	17.	139	129	12	8	44	23.	131	224	109	38
18	24.	118	110	11	7	45	30.	124	226	101	34
19	May 1.	92	136	17	10	46	Novem. 6.	136	183	68	27
20	8.	116	103	13	11	47	13.	125	162	41	20
21	15.	128	94	13	8	48	20.	121	145	28	11
22	22.	113	132	14	9	49	27.	143	123	22	13
23	29.	94	98	9	7	50	Decem. 4.	155	160	45	17
24	June 5.	129	112	16	8	51	11.	135	137	38	20
25	12.	127	112	19	14	52	18.	136	132	28	15
26	19.	121	119	15	10	53	25.	134	135	38	19
27	26.	132	126	24	16						

The Totals {
Chriſtened — — 6614
Buried — — 7920
Whereof of the Plague 2124
}

∗Bell's London's Remembrancer.

Table B.3: A Table of the Christenings and Mortality For the Year 1605 and 1606.∗

> rope that the English are fools and madmen. Fools, because they give their children the smallpox to prevent their catching it; and madmen, because they want only communicate a certain and dreadful distemper to their children, merely to prevent an uncertain evil.[138] (Voltaire)

It would, however, be unfair to disregard this opposition as gross errors as in fact, the scientific debate of the time to decide whether to inoculate the small pox or not was very much alike what one could observe today. Most prominently (for us at least) is the very public debate between Daniel Bernoulli and Jean Le Rond d'Alembert. While both agreed on principle on the usefulness of variolation, they disagreed fervently on the scientific reasoning (and maybe, at least in part, because of personal rivalry). We will not expose here the reasons of the disagreement as they mostly pertain on the philosophy of science and the use of statistics to guide it, D'Alembert also argues on the personal benefit using arguments that readers familiar with macroeconomics will recognize as considerations of *utility* or risk convexity for the more econometrically inclined. It is, however, a subject still relevant today and we refer the curious reader to Colombo and Diamanti (2015) for more details. However the argument brought forward by Bernoulli in front of the Académie des Sciences de Paris in 1760 further proved the importance of studying *lifetimes* in order to take decisions.

After the question of the variolation was made the central question of the scientific talks of the Académie des Sciences by Charles Marie de La Condamine in 1754, Bernoulli proposed a dynamical model of the smallpox with the goal of comparing the outcomes in terms of mean survival times, i.e. life expectation, between the variolated and non-variolated populations. In Bernoulli's model, death can happen randomly either because of smallpox or due to other factors, it is, however, possible to survive smallpox such that immunity is acquired, leaving only the other factors of death as possible causes. By viewing variolation as a form of induced immunity, Bernoulli is able to model the variolated and unvariolated populations using the exact same mechanics but with different initial conditions. Using the modern notation of differential equations, we can denote by $s(t)$ the fraction of the population that is infected and survives at time t and $1 - s(t)$ the fraction that is infected and dies.[139] We then denote by $\lambda(t)$ the instantaneous rate of infection to smallpox, while $\mu(t)$ is the instantaneous of *death by other causes*. If we take $u(t)$ the probability of a newborn to still be alive and susceptible at age t and $w(t)$ the probability to be immune and alive,[140] then both follow the ODE

138: "On dit doucement, dans l'Europe chrétienne, que les Anglais sont des fous et des enragés: des fous, parcequ'ils donnent la petite vérole à leurs enfants, pour les empêcher de l'avoir, des enragés, parce qu'ils communiquent de gaieté de coeur à ces enfants une maladie certaine et affreuse, dans la vue de prévenir un mal incertain."

139: That is $s(t)$ is the survival rate of smallpox and $1 - s(t)$ is the case fatality rate.

140: That is, the subsurvivals of an individual at age 0.

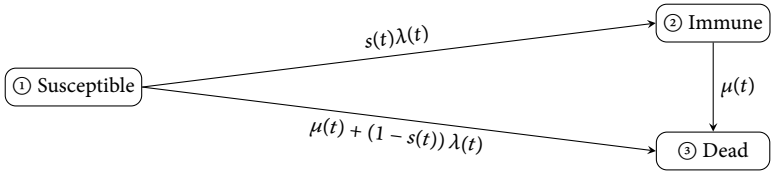


FIGURE B.5: Bernoulli's epidemiological model of the smallpox.

$$\frac{\mathrm{d}u}{\mathrm{d}t} = -\big(\lambda(t) + \mu(t)\big)\,u(t) \qquad u(0) = 1$$
$$\frac{\mathrm{d}w}{\mathrm{d}t} = s(t)\lambda(t)u(t) - \mu(t)w(t) \qquad w(0) = 0$$

This population dynamics model proposed by Bernoulli, represented graphically in fig. B.5, can be seen today as multi-state survival model and therefore represents one of the earliest examples of survival analysis. More surprisingly,[141] Bernoulli was able to show that the previous equation admits the solution

$$u(t) = \exp(-\Lambda(t) - M(t))$$
$$w(t) = e^{-M(t)} \int_0^t s(y)\lambda(y)e^{-\Lambda(y)}\,\mathrm{d}y$$

where Λ and M are what would now be known in the literature as the cumulative hazards

$$\Lambda(t) = \int_0^t \lambda(y)\,\mathrm{d}y$$
$$M(t) = \int_0^t \mu(y)\,\mathrm{d}y$$

We skip here the intermediate details, which are given in great details in Dietz and Heesterbeek (2002) or Colombo and Diamanti (2015) for the modern interpretation or simply in Bernoulli (1766). If we denote by $S(t)$ the probability of surviving at age t, then we have

$$S(t) = u(t) + w(t) = S_0(t)\left(e^{-\Lambda(t)} + \int_0^t s(y)\lambda(y)e^{-\Lambda(y)}\,\mathrm{d}y\right)$$

with

$$S_0(t) = e^{-M(t)},$$

the baseline survival function of the population without smallpox. If we then introduce $x(t) = u(t)/S(t)$ the proportion of at risk individuals at age t we can then derive the differential equation

$$\frac{\mathrm{d}x}{\mathrm{d}t} = -\lambda(t)x(t)\,(1 - x(t) - s(t)x(t)) \qquad x(0) = 1,$$

141: Or maybe unsurprisingly given Bernoulli's pedigree and accomplishments.

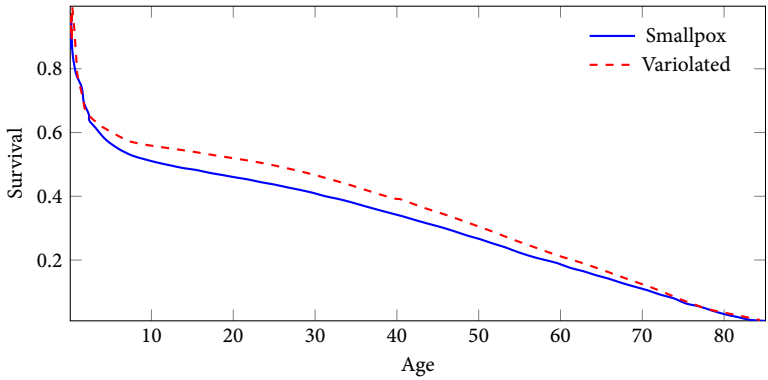


FIGURE B.6: Increased survival of the smallpox variolated population compared to the population with smallpox present.

which does not involve the mortality $\mu(t)$. Quite luckily, this differential equation was already known at the time to be solvable through the work of the *other* Bernoulli, Jacob, Daniel's uncle..[142] It is possible to show that

$$x(t) = \frac{e^{-\Lambda(t)}}{e^{-\Lambda(t)} + \int_0^t s(t)\lambda(t)e^{-\Lambda(t)}\,\mathrm{d}y}.$$

All that is left for the model to be complete is then to reconstruct the missing quantities from real-world data in order to obtain the desired estimate of the survival function of the population with smallpox present compared to a population without smallpox, that is a hypothetically variolated at birth population. Fortunately for us, and Bernoulli, Edmond Halley was able in 1693 to establish a life table for the deaths from smallpox in the city of Breslau much akin to that of Graunt (Halley [1693]), reproduced here in table B.4. From that table Bernoulli was able to infer the missing variables in order to derive the survival functions of both the unvariolated and variolated populations in fig. B.6. After numerically[143] integrating the aforementioned survival functions, Bernoulli was then able to deduce the life expectancy with smallpox to be 26.57 years compared to 29.65 years without smallpox; a more than 3-year increase in life expectancy.[144] However as, already at the time, most opponents of variolation accurately objected; variolation was not a perfectly safe[145] procedure. In order to account for that fact, Bernoulli considered a probability p of dying from the act, and showed that in order for the life expectancy gain to disappear we need $p \geq 0.11$. As Bernoulli estimated that the true risk of variolation was $p = 0.01$, his position was clear:

> I simply hope that, in a question that so closely regards the wellbeing of the human race, no decision will be taken without considering all the information that a modest analysis and calculation can provide. (Daniel Bernoulli)

142: Who introduced *Bernoulli's equations* as well as the law of large numbers. Not to be confused with Johann Bernoulli, the *other other* Bernoulli, known for his contributions to infinitesimal calculus as well as educating Leonhard Euler. Work on cloning the Bernoulli's family may be a valid future research direction.

143: And manually!

144: The reader may be shocked by such a low life expectancy, but one has to remember that people *did* live old, infantile mortality was just at the time incredibly high.

145: Variolation, contrary to vaccination, involves the exposition to a live virus but, hopefully, in a non-lethal quantity.

The method employed here by Bernoulli show a different type of question: "How do the disease react through time?" and is therefore meant as a tool for *decision* by making the interactions and mechanisms explicit such that it is possible to study the impact of a change and therefore of an decision. Bernoulli's approach is surprisingly modern and can be considered a precursor to the Susceptible, Infectious, or Recovered (SIR) model still in use today for the modelling of SARS-CoV-2 and policy design during the COVID-19 pandemic (Y.-C. Chen et al. [2020]; Cooper, Mondal, and Antonopoulos [2020]).

Ages par années.	Survivans selon M. Halley.	N'ayant pas eu la pet. vérole.	Ayant eu la pet. vérole.	Prenant la pet. vérole pendant ch. année.	Morts de la pet. vérole pendant caq. ann.	Somme des morts de la pet. vérole.	Morts par d'autres maladies pend. chaq. année.
0	1300	1300	0				
1	1000	896	104	137	17.1	17.1	283
2	855	685	170	99	12.4	29.5	133
3	798	571	227	78	9.7	39.2	47
4	760	485	275	66	8.3	47.5	30
5	732	416	316	56	7.0	54.5	21
6	710	359	351	48	6.0	60.5	16
7	692	311	381	42	5.2	65.7	12.8
8	680	272	408	36	4.5	70.2	7.5
9	670	237	433	32	4.0	74.2	6
10	661	208	453	28	3.5	77.7	5.5
11	653	182	471	24.4	3.0	80.7	5
12	646	160	486	21.4	2.7	83.4	4.3
13	640	140	500	18.7	2.3	85.7	3.7
14	634	123	511	16.6	2.1	87.8	3.9
15	628	108	520	14.4	1.8	89.6	4.2
16	622	94	528	12.2	1.6	91.2	4.4
17	616	83	533	11.0	1.4	92.6	4.6
18	610	72	538	9.7	1.2	93.8	4.8
19	604	63	541	8.4	1.0	94.8	5
20	598	56	542	7.4	0.9	95.7	5.1
21	592	48.5	543	6.5	0.8	96.5	5.2
22	586	42.5	543	5.6	0.7	97.2	5.3
23	579	37	542	5.0	0.6	97.8	6.4
24	572	32.4	540	4.4	0.5	98.3	6.5

Table B.4: Halley's table as reproduced and extended by Bernoulli.

Introduction - Français | C

La tâche étudiée dans cette thèse est celle de la prédiction de la mort d'individus à partir de leurs caractéristiques. Cette tâche, généralement connue sous le nom d'*analyse de survie*, et à l'origine intrinsèquement liée à celle de l'épidémiologie, a une riche histoire mathématique et a évolué en suivant les progrès constants des statistiques. La modélisation du décès d'un individu est restée pendant des siècles l'un des problèmes phares de la recherche médicale et de la biostatistique pour la simple raison que comprendre la cause du décès est un premier pas vers la prévention de ce décès. Ainsi, les statisticiens ont pu fournir aux chercheurs médicaux les outils mathématiques nécessaires pour répondre aux questions médicales de manière scientifique, comparer la survie de sous-populations ou quantifier la certitude de leurs hypothèses. Cette thèse comportera de nombreux exemples médicaux de ce type, non seulement parce qu'il s'agit d'une application intéressante et utile, mais aussi parce que les chercheurs médicaux ont généreusement offert à la communauté scientifique un grand nombre de données ouvertes sur lesquelles il est possible de tester de nouvelles approches. Ce mémoire ne porte toutefois pas sur la médecine, mais sur la finance et il ne s'agira pas ici de comprendre les causes et les effets ou de prouver statistiquement des affirmations, mais uniquement de *prédire* la mort. Cette thèse, bien qu'en grande partie théorique, a pour objectif principal de répondre aux besoins pratiques de BNP Paribas et en particulier du département Portfolio Management de la branche CIB, dont le rôle principal est de réduire l'exposition de la banque au risque de crédit en gérant activement celui-ci. Afin de gérer ce risque, nous devons prédire avec précision les événements potentiels, souvent en utilisant des données très peu structurées et volumineuses, ce qui motive naturellement une étude rigoureuse du risque de crédit dans le cadre de l'analyse de survie et de l'apprentissage automatique.

C.1 Vie et mort d'une entreprise

Nous avons surtout décrit l'étude de la mort au travers de l'étude de la mort des individus, mais celle-ci n'est pas réservée aux êtres vivants. Même dans le langage courant, il n'est pas rare de désigner la défaillance catastrophique d'un objet comme sa *mort* ; "Mon téléphone est mort !". Si la

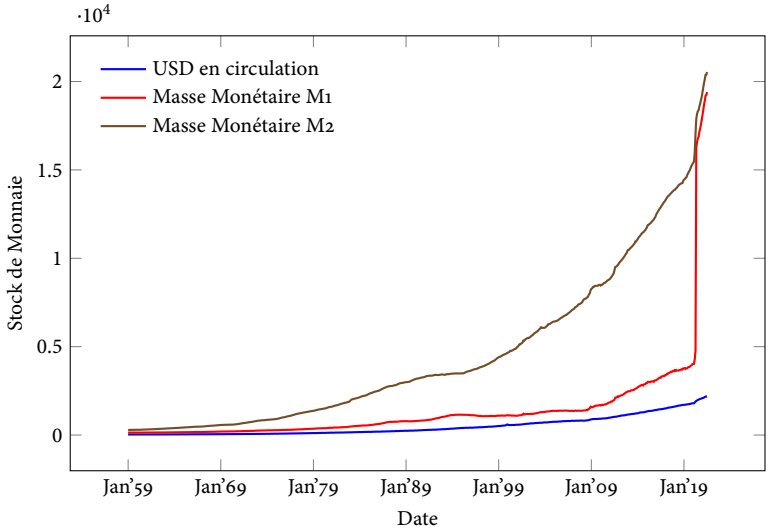


FIGURE C.1 : Monnaie *sonnante et trébuchante vs* masse monétaire.

mort peut également frapper des objets inanimés, il semble alors judicieux, ou du moins utile de la prévoir avant qu'elle ne survienne peut-être par exemple pour mettre de côté de l'argent[147] pour un nouveau téléphone avant que le précédent ne meure subitement. Dans ce cas précis, à savoir la prédiction de la défaillance d'un élément mécanique, prévoir la défaillance avant qu'elle ne se produise permet de programmer la maintenance à l'avance afin d'être parcimonieux sur les coûteux contrôles et opérations de maintenance. Ce problème, connu dans la littérature sous le nom de maintenance prédictive (voir Zonta et al. [2020] ; Bousdekis et al. [2019] ; Ran et al. [2019], pour un aperçu), peut naturellement être traité comme un problème d'analyse de survie (see e.g. C. Chen et al. [2020]).[148] Une interprétation similaire apparaît naturellement dans le monde financier à travers le concept de *credits*. Depuis l'avènement de la banque, et plus tard des réserves fractionnaires, les crédits en sont venus à représenter la majorité des actifs monétaires en circulation[149] car elle libère des liquidités qui peuvent être ensuite réutilisées dans l'économie de façon plus productive. Ce phénomène s'est fortement accéléré ces dernières années, car les gens ont fini par accepter la dissociation entre les concepts de monnaie, i.e. moyen d'échange, et de physicalité. Dans figure C.1, nous représentons le stock monétaire du dollar américain, où le stock M1 englobe la monnaie hors du Trésor américain, les dépôts dans les banques commerciales et autres dépôts vérifiables et le stock M2 est constitué du stock M1 plus les dépôts d'épargne et les soldes des fonds monétaires de détail. Bien qu'il ne s'agisse pas d'un problème en soi, le fait que la majeure partie du bilan des entreprises ou même des individus[150] consiste désormais en des prêts

147 : Dans un monde parfait du moins. Ou de manière plus réaliste si l'on est responsable d'une flotte de milliers de téléphones d'entreprise.

148 : Il est également possible de le voir comme un problème de bandit (Ruiz-Hernández, Pinar-Pérez et Delgado-Gómez [2020] ; Fouché, Komiyama et Böhm [2019]).

149 : On pourrait dire que la monnaie elle-même ne représente qu'un *bon-pour-avoir* ou un crédit sous forme physique.

150 : L'argent que j'ai aujourd'hui sur mon compte bancaire est en réalité un prêt, dont j'espère que ma banque ne fera pas défaut.

qui comportent un risque de contrepartie signifie que la richesse doit être traitée comme une variable aléatoire. Un *défaut* est alors la mort d'un prêt.

C.1.1 Défauts et contagions

Si, du point de vue de l'emprunteur, les prêts peuvent effectivement être considérés comme de l'argent avec décote puisqu'il reçoit des espèces sonnantes et trébuchantes, il n'en va pas du tout de même pour le prêteur. Du point de vue du prêteur, les prêts comportent une part importante d'incertitude ou de risque appelé *risque de contrepartie*[151] : l'emprunteur peut très bien ne jamais rembourser son prêt. La valeur d'un prêt, c'est-à-dire la somme d'argent que le prêt rapportera,[152] est donc une quantité aléatoire. Si l'emprunteur rembourse intégralement son prêt, alors la valeur réalisée sera le prêt plus les intérêts tandis que si, pour une raison quelconque, le prêt n'est pas remboursé, la valeur ne sera alors égale qu'au principal et aux intérêts remboursés jusqu'au moment du défaut. Il est clair que le gain réalisé est par nature stochastique et dépend de l'événement aléatoire "a remboursé son emprunt" dont la probabilité est primordiale. À partir de cette observation, il est possible de définir la *juste valeur* d'un prêt, que pour des raisons de simplicité nous supposerons ici être le bénéfice attendu,[153] et à partir de cette définition de la juste valeur, il est possible de trouver le taux auquel un prêt devrait être émis. Il y a cependant un inconvénient important à la remarque précédente : en raisonnant en termes de valeur espérée, nous masquons le fait que les individus ne disposent que d'une quantité finie d'argent et ne peuvent donc survivre qu'à une quantité finie de pertes. Ce ne serait pas un problème si toutes les entités disposaient de plus de liquidités que de prêts en cours, mais nous avons vu précédemment dans figure C.1 que, pour de bonnes raisons, la quantité d'argent liée à des prêts dépasse largement la quantité d'argent détenue en propre. Il est donc possible, et même garanti après suffisamment de temps (voir Embrechts, Klüppelberg et Mikosch [1997], pour la théorie de la ruine, e.g. figure C.2), qu'un événement extrême [154] se produise et soit plus dommageable à ce que le prêteur peut supporter. Si de tels événements catastrophiques sont en théorie rares, leur impact peut avoir des répercussions catastrophiques pour les mêmes raisons que les virus ont un impact important sur la population globale. Les individus, ou dans ce cas les entreprises n'existent pas en vase clos et interagissent les uns avec les autres. Si, dans le cas de la peste ou de la variole, cette interaction peut entraîner la propagation d'une charge virale, dans le cas des entreprises, elle se traduit par la propagation de pertes de crédit. Une entreprise subissant une perte de crédit extrême au point d'entraîner son propre défaut est un événement incroyablement rare pris individuellement, mais une défaillance systématique d'une multitude de prêts et d'entreprises est un

151 : Dans le cas le plus général. Nous faisons ici l'amalgame par abus entre le risque de crédit et le risque de contrepartie.

152 : Eventuellement actualisée au taux sans risque, mais nous ignorerons ici toutes les préoccupations d'ordre purement financier.

153 : En pratique, tous les produits financiers sont évalués comme la *valeur attendue* de leur gain, mais la valeur attendue n'est pas nécessairement prise sous la mesure de probabilité naturelle mais souvent sous une probabilité différente appelée mesure risque-neutre. Cela sort du cadre de cette thèse mais les lecteurs curieux peuvent se référer à Shreve (2004).

154 : C'est-à-dire un événement situé profondément dans la queue de distribution des pertes.

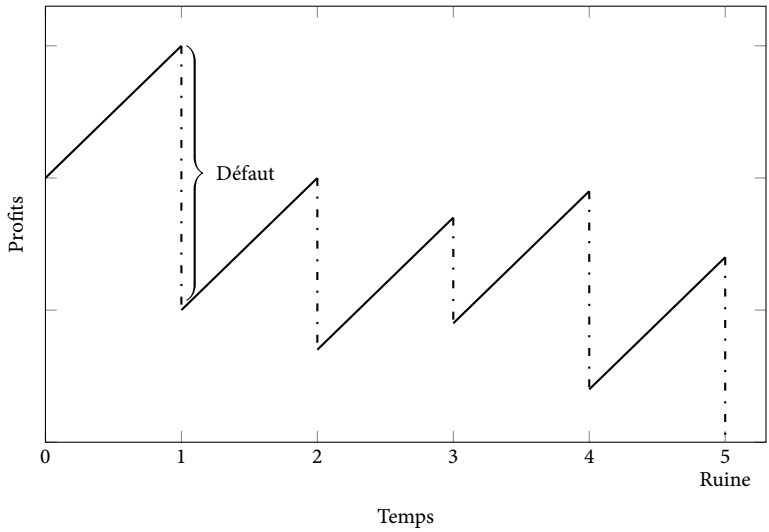


FIGURE C.2 : Actifs d'une entreprise provenant de prêts vus comme un processus de Poisson composé.

événement beaucoup plus probable conditionné à ce premier défaut, ou *patient zéro*. Une entreprise subissant une perte suffisante pour provoquer sa propre défaillance est, par définition, incapable d'honorer ses propres prêts ce qui, selon la taille de l'entité qui s'effondre, peut provoquer la défaillance en cascade d'autres entreprises. Au fur et à mesure que d'autres entreprises ne remboursent pas leurs prêts, le phénomène se propage dans le complexe maillage des relations financières et peut, dans le pire des cas, provoquer une crise financière d'ampleur globale. De tels événements ont conduit à la crise hypothécaire de 2008 où la corrélation entre les entités, et par conséquent le risque de propagation de ce virus financier, a été sous-estimée.[155] Étant donné les parallèles entre les virus biologiques et le risque de crédit, il n'est donc pas étonnant que l'étude des pandémies ait inspiré le traitement des crises financières.

C.1.2 Vaccination réglementaire

Après la crise de 2008, des mesures sans précédent ont été mises en œuvre afin de vacciner le monde financier contre le risque de crédit. Alors que les vaccins reposent sur l'administration d'une quantité minimale d'antigènes afin de survivre à une charge virale normale,[156] les réglementations en matière de crédit telles que BÂLE II à BÂLE IV (Basel Committee [2019]), s'appuient sur la présence d'un montant tampon minimum de liquidités pour survivre aux situations extraordinaires ou "inattendues" [157] comme représenté dans figure C.3, afin d'assurer des pertes sans défaillance et donc sans contamination des autres contreparties. Afin d'assurer la résilience

155 : Les lecteurs de cette thèse peuvent également avoir le cas plus récent d'Evergrande comme exemple en fonction de la façon dont les événements se sont déroulés.

156 : S'il vous plaît, ne prenez pas mes explications comme autre chose qu'une métaphore, se référer à A. J. Pollard et Bijker (2021).

157 : Inattendues au sens commercial, pas au sens mathématique. L'abus de termes mathématiques est une tradition de longue date dans le monde financier, juste derrière celle d'inventer des nouvelles lettres grecques.

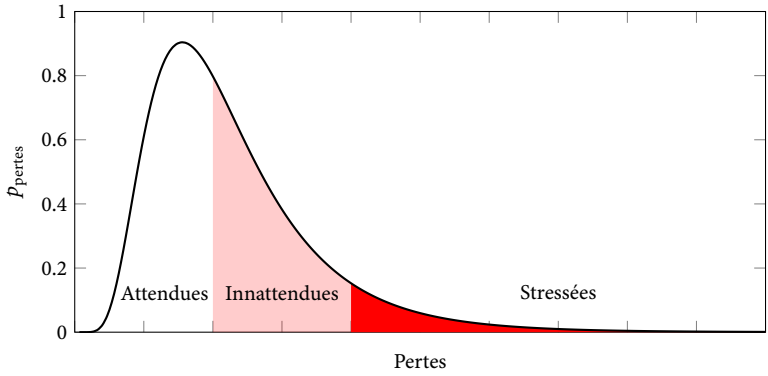


FIGURE C.3 : Pertes attendues, inattendues et stressées

des acteurs financiers aux pertes extrêmes, ceux-ci sont tenus de détenir un montant minimum de capital pour compenser les actifs risqués désignés ici par RWA tel que

$$\frac{\text{Capital}}{\text{RWA}} \geq 0.08.$$

Afin d'harmoniser les approches, et surtout d'éviter les calculs douteux du RWA et donc de graves sous-estimations[158] du capital requis, le BIS fournit une approche standard pour la détermination des quantités requises. Pour les prêts aux entreprises par exemple, le dispositif de Bâle III impose

$$\text{Capital} = \text{LGD} \times \left(\Phi\left(\sqrt{\frac{1}{1-R}}\Phi^{-1}(p) + \sqrt{\frac{R}{1-R}}\Phi^{-1}(0.999) \right) - p \right),$$

$$\text{RWA} = \frac{\text{Capital} \times \text{EAD}}{0.08},$$

où R est un facteur de corrélation défini par

$$R = A\left(0.12 \times \frac{1-e^{-50p}}{1-e^{-50}} + 0.24 \times \left(1 - \frac{1-e^{-50p}}{1-e^{-50}}\right)\right),$$

et $A \in \{1, 1.25\}$ est un facteur dépendant de la taille de l'institution,[159] Φ est la fonction de répartition de la loi normale standard et p est la probabilité de défaut. Comme la probabilité de défaut est la seule quantité qui n'est pas explicitement donnée, et le cœur des valeurs précédentes comme l'illustre rapidement figure C.4, son estimation joue un rôle central dans la stratégie commerciale des organisations financières. En effet, si chez l'homme la vaccination ne présente pas d'inconvénient majeur autre qu'un bras potentiellement douloureux ou un syndrome grippal, ce n'est pas le cas chez les acteurs financiers. Se préparer à un défaut catastrophique et suivre les réglementations implique de geler une quantité importante d'argent

158 : Volontaires ou non.

159 : Afin de représenter le degré de centralité et de contact aux autres de l'institution.

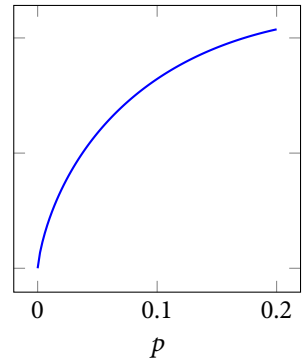


FIGURE C.4 : Capital requis en fonction de la probabilité de défaut.

160 : Les rendements en finance sont directement liés au risque. Les actifs sans risque n'offrent donc aucun rendement, voire des rendements négatifs.

dans des actifs sans risque[160] et donc un coût d'opportunité important avec, dans le pire des cas, des pertes économiques réelles. Si le régulateur fournit des lignes directrices sur l'estimation de p sur la base de la notation donnée par les agences externes, il offre également certaines libertés et permet à cette quantité clé d'être modélisée en interne en utilisant l'approche dite IRBA.

La plupart des approches en risque de crédit traitent le problème comme un problème de classification, prédisant soit le défaut, soit l'absence de défaut. Cependant, cette approche présente de sérieux inconvénients, car elle repose sur la discrétisation du temps et sur la décision arbitraire de classer quelque chose comme défaut ou non-défaut sur la base d'un seul horizon temporel τ. Dans cette approche, une entreprise qui fait défaut après $\tau+1$ jours est considérée comme un bon payeur, une décision somme toute assez discutable. Pour cette raison, et parce que l'on rencontre exactement les mêmes problèmes dans le domaine médical, nous adopterons ici une approche différente : au lieu de prédire un événement binaire, "défaut" *vs* "non-défaut" [161] nous prédirons le *temps jusqu'au défaut*, ou plus généralement le *temps jusqu'à l'événement*, en adoptant le point de vue que toutes les entreprises finissent par faire défaut et que nous n'avons qu'éventuellement pas pu l'observer.

161 : ou "mort" *vs* "non-mort" ou "échec" *vs* "non échec".

C.2 Temps jusqu'a l'événement

La prédiction de l'occurrence d'un événement peut être abordée sous de multiples angles, le plus simple étant de traiter le problème comme un problème de classification binaire. Après avoir choisi un seuil temporel τ, il est possible de reformuler la plupart des problèmes impliquant la survie d'un individu, vivant ou non, comme le problème de classification "l'événement s'est-il produit avant τ ou non ?". Cette approche simpliste s'avère être bien adaptée à de nombreux problèmes où l'acte de choisir un seuil est en soi naturel, par exemple pour un emprunt spécifique d'une durée prédéterminée τ, mais elle est souvent en pratique très insuffisante. La plupart des problèmes ne peuvent pas être binarisés naturellement ; pour une ligne de crédit perpétuelle, si l'on fixe le seuil à τ, cela signifie-t-il que les clients qui font défaut à $\tau + 1$ sont de bons clients ? Dans le cadre médical, si le but est de comparer deux traitements, où doit-on fixer τ ? Pour un grand τ on observe potentiellement que des décès naturels ou même rien du tout si l'étude est trop courte et pour un petit τ on court le risque de ne pas avoir attendu assez longtemps pour observer quoi que ce soit. De plus, en traitant le problème comme une tâche de classification, on est rapidement confronté à des problèmes de déséquilibres des classes puisque, heureusement, la plupart des clients ne font pas défaut et la plupart

FIGURE C.5 : *Memento mori.* Pour Benjamin Franklin, "Rien ne peut être considéré comme certain, sauf la mort et les impôts". Nous demandons au lecteur d'ajouter les défauts à cette liste pour le reste de cette thèse.

des gens ne meurent pas, ce qui résulte en des instances de classification difficiles.

Pour ces raisons et bien d'autres, il est donc plus naturel de traiter le problème de la prédiction d'un événement comme le problème de la prédiction d'un *temps jusqu'à un événement*, c'est-à-dire l'apprentissage de la distribution des moments à partir d'aujourd'hui où l'événement d'intérêt se produit. Si nous désignons le temps jusqu'à l'événement par $Y \in \mathbb{R}_+$, suivant la notation habituelle de la régression, que nous supposons par convention être une variable aléatoire positive, nous souhaitons alors estimer la distribution de Y à travers l'une des multiples quantités qui la définissent. Comme le sujet s'appelle *analyse de survie* et que nous nous intéressons aux décès, aux défaillances et aux défauts et non à la période où rien ne se passe, nous choisissons généralement d'étudier la fonction de survie $S(t) = \mathbb{P}(Y > t)$ au lieu de la fonction de répartition $F(t) = 1 - S(t)$. La fonction de survie joue ici le rôle d'extension naturelle de l'approche précédente par seuillage puisque chaque instance $S(\tau)$ représente un problème de classification binaire, on résout donc ici *toutes* ces instances en même temps. Bien entendu, d'autres quantités d'intérêt peuvent être modélisées en fonction des particularités du problème, comme décrit plus tard dans figure 3.1. En particulier, dans le cas où Y admet une densité $p(t)$,[162] nous pouvons définir le risque instantané

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{\mathbb{P}\left(Y \in [t, t + \Delta t] \mid Y > t\right)}{\Delta t},$$

qui se rapporte à la survie naturellement par la relation

$$\lambda(t) = \frac{p(t)}{S(t)}.$$

De même, le risque intégré ou cumulatif $\Lambda(t) = \int_0^t \lambda(t)\,\mathrm{d}t$ est souvent étudié en raison de la relation

$$S(t) = \exp(-\Lambda(t)),$$

qui découle trivialement de la définition de λ. Notez que toutes ces quantités définissent de manière unique la loi de Y et peuvent donc être utilisées de manière interchangeable.

Bien que la formulation temps jusqu'à l'événement soit particulièrement bien adaptée à notre problème, nous devons malheureusement encore composer avec un horizon temporel. Non seulement nos observations s'arrêtent nécessairement à l'instant présent, ou du moins au moment où nous avons cessé de collecter des données,[163] mais certaines observations ne sont pas observées pour des raisons indépendantes de notre volonté, comme le fait qu'un patient abandonne une étude ou qu'une entreprise

162 : La notation inhabituelle de p au lieu de celle plus commune de f pour la densité sera utilisée dans cette thèse.

163 : Par exemple, à la fin d'une étude médicale.

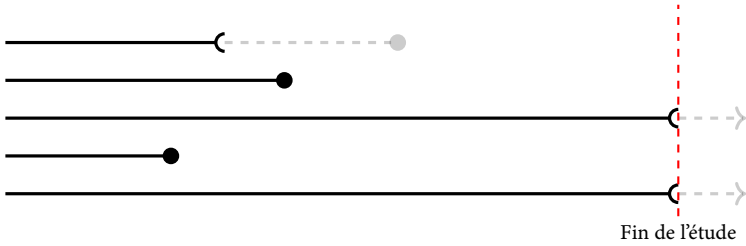


Figure C.6 : Données de survie censurées à droite.

fusionne avec une autre et de fait "disparaisse". En réalité, nous pouvons donc rarement observer Y, la véritable variable d'intérêt, et nous n'observons qu'un certain temps T que nous qualifierons de *censuré à droite* tel que

$$T = \min(Y, C)$$

où $C \in \mathbb{R}_+$ est une variable aléatoire de nuisance, jouant un rôle symétrique à Y et appelée ici *variable de censure*, qui englobe toutes les raisons pour lesquelles notre variable d'intérêt peut être inobservée comme l'écoulement du temps, la fin de l'étude, une observation retirée de l'ensemble de données etc. Nous supposons également que nous savons si le temps que nous observons est censuré ou non grâce à travers l'*indicateur de censure* δ défini par

$$\delta = \mathbb{1}_{Y \leq C} = \begin{cases} 1 & \text{si } Y \leq C \\ 0 & \text{sinon.} \end{cases},$$

car il n'y aurait sans ce dernier aucun espoir d'estimer une quelconque quantité utile. Notre ensemble de données, représenté dans figure C.6, est donc constitué d'observations du couple (T, δ) au lieu de Y.

Cette quantité, paramètre phare de l'analyse de survie,[164] comme cela a été largement étudié dans la littérature statistique : (voir Fleming et Harrington [1991] ; ou Gill [1994], pour un excellent aperçu des méthodes utilisées pour obtenir les estimateurs) qui s'est concentré sur les propriétés asymptotiques des divers estimateurs de S ou Λ.

164 : Il est possible de définir également la censure à gauche ainsi que la troncature gauche/droite. La plupart des résultats présentés ici peuvent être adaptés aux cas plus généraux sans difficulté autre que technique.

C.2.1 Estimateurs de la survie

Le domaine de l'analyse de survie étant trop vaste pour être résumé ici, nous ne donnerons qu'une brève présentation de l'estimateur clé de la survie qui servira d'inspiration aux chapitres suivants. Une étude rigoureuse de la littérature est reportée aux chapitres pertinents. Cependant, si le lecteur s'intéresse à l'analyse de survie, nous recommandons Klein et Moeschberger (2003) ; D. R. Cox et Oakes (1984) pour un aperçu général

ainsi que le cours mentionné précédemment de Gill (1994) qui motive la formulation produit-intégral.

Dans le cas où l'objet d'intérêt est la survie S,[165] si nous avions observé la variable réelle d'intérêt Y, nous pourrions facilement estimer S par

$$S_n(t) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{Y_i > t},$$

qui, par Glivenko-Cantelli, converge uniformément vers la vraie survie. Ici, nous n'observons que (T, δ) et l'estimateur correspondant serait

$$\bar{S}_n(t) = \frac{\sum_{i=1}^{n} \mathbb{1}_{T_i > t, \delta_i = 1}}{\sum_{i=1}^{n} \mathbb{1}_{\delta_i = 1}},$$

qui est biaisé et ne converge pas vers la valeur d'intérêt.

Cependant, après avoir discrétisé le temps à chaque observation T_i, nous pouvons appliquer la formule de Bayes après avoir remarqué que localement, à l'intérieur d'un intervalle $[T_{[i]}, T_{[i+1]}]$ où $T_{[k]}$ est utilisé pour signifier "la k-ième plus grande valeur de (T_i)" (en ignorant les égalités pour simplifier), nous pouvons écrire la probabilité conditionnelle qu'un événement se produise dans cet intervalle étant donné que rien ne s'est produit jusqu'à présent, comme si aucune censure n'était présente. Autrement dit, si nous désignons par m_i le nombre d'événements dans le i-ième intervalle, n_i le nombre d'*individus à risque*, c'est-à-dire vivants et non censurés et c_i le nombre d'individus censurés au début de l'intervalle i.e. à T_i alors

$$\mathbb{P}\left(Y \leq T_{[i+1]} \mid Y > T_{[i]}\right) = \frac{m_i}{n_i - c_i}.$$

Nous pouvons donc construire itérativement un estimateur de S de la forme

$$\hat{S}_n(t) = \prod_{i | T_i \leq t} \left(1 - \frac{m_i}{n_i - c_i}\right),$$

ou réécris sous plusieurs formes équivalentes différentes

$$\begin{aligned}\hat{S}_n(t) &= \prod_{i=1}^{n} \left(1 - \frac{\delta_{[i]}}{n - i + 1}\right)^{\mathbb{1}_{T_{[i]} \leq t}} \\ &= \prod_{\substack{i=1,\dots,n \\ T_{[i]} \leq t}} \left(\frac{n-i}{n-i+1}\right)^{\delta_{[i]}}.\end{aligned}$$

On peut montrer que cet estimateur, souvent appelé estimateur de Kaplan-Meier (Kaplan et Meier [1958]), est consistant. Des résultats similaires

165 : Ce qui est souvent le cas. Dans le cadre médical, nous sommes par exemple intéressés par la comparaison entre S_1 et S_2, les fonctions de survie de deux traitements concurrents, afin de prouver une hypothèse du type $S_1 > S_2$ ou du moins $S_1 \neq S_2$

peuvent être obtenus pour le risque cumulatif Λ avec l'estimateur de Nelson-Aalen (Nelson [1969] ; Aalen [1978])

$$\hat{\Lambda}_n(t) = \sum_{i=i}^{n} \frac{\delta_i \mathbb{1}_{T_i \leq t}}{\sum_{j=1}^{n} \mathbb{1}_{T_j > T_i}}.$$

Bien que nous ayons ici ignoré les covariables X et donc le conditionnement sur X, ces estimateurs nous intéressent particulièrement en raison de la facilité d'introduction de ce conditionnement : en introduisant une moyenne locale autour de X, par exemple par le biais de noyaux, on peut obtenir des estimateurs conditionnels à $X = x$ de la forme

$$\tilde{\Lambda}_n(t \mid X = x) = \sum_{i=i}^{n} \frac{\delta_i \mathbb{1}_{T_i \leq t} K(x - X_i)}{\sum_{j=1}^{n} \mathbb{1}_{T_j > T_i} K(x - X_j)},$$

où K est généralement une fonction de densité de probabilité[166] symétrique autour de 0.

166 : Ou noyaux.

C.2.2 Modèles paramétriques et semi-paramétriques

Une observation surprenante qui débloque un grand nombre de techniques déjà existantes dans le cadre non censuré réside dans la décomposition conditionnelle de la vraisemblance des observations. Si nous supposons pour l'instant que Y et C sont indépendants, nous pouvons écrire la vraisemblance de l'observation i comme suit

$$\begin{aligned}
&\mathbb{P}\left(T \in [T_i, T_i + \mathrm{d}t], \delta = \delta_i \mid \theta\right) \\
&\quad = \mathbb{P}\left(T \in [T_i, T_i + \mathrm{d}t], \delta = 1 \mid \theta\right)^{\delta_i} \\
&\qquad \times \mathbb{P}\left(T \in [T_i, T_i + \mathrm{d}t], \delta = 0 \mid \theta\right)^{1-\delta_i} \\
&\quad = \mathbb{P}\left(Y \in [T_i, T_i + \mathrm{d}t], C \geq T \mid \theta\right)^{\delta_i} \\
&\qquad \times \mathbb{P}\left(C \in [T_i, T_i + \mathrm{d}t], C < T \mid \theta\right)^{1-\delta_i} \\
&\quad = \left(p\left(T_i \mid \theta\right) S_C(T_i-)\right)^{\delta_i} \left(p_C(T_i) S(T_i \mid \theta)\right)^{1-\delta_i}, \qquad \text{(C.1)}
\end{aligned}$$

où $\theta \in \Theta$ est ici le paramètre décrivant la famille d'intérêt, c'est-à-dire la distribution de la survie et p_C, S_C sont la densité et la survie de la variable de censure. La quantité précédente de équation (C.1) implique à la fois les objets d'intérêt p et S mais aussi les quantités de nuisance p_C, S_C ce qui semble à première vue être un problème. Cependant, comme C est précisément une variable de *nuisance* et donc non pertinente pour nous, il n'est pas utile de la modéliser et elle ne fait donc apparaître θ d'aucune façon . Ainsi, après avoir ignoré ces quantités en les traitant comme des constantes, nous pouvons écrire la vraisemblance comme suit

$$\mathcal{L} \propto \prod_{i=1}^{n} p(T_i \mid \theta)^{\delta_i} S(T_i \mid \theta)^{1-\delta_i}. \qquad \text{(C.2)}$$

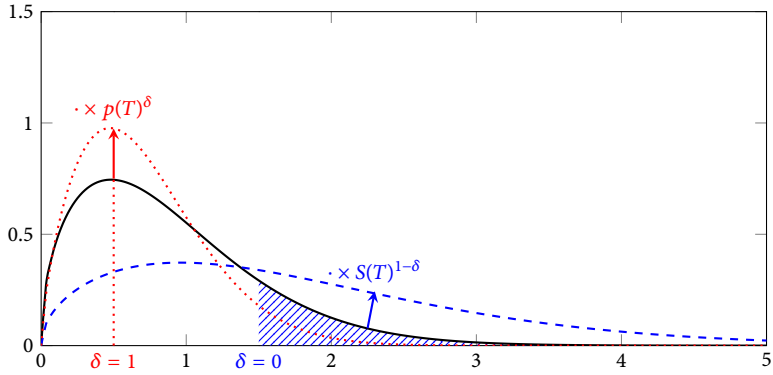


FIGURE C.7 : Contribution individuelle des observations censurées et non censurées.

La quantité de équation (C.2) est souvent appelée la *vraisemblance partielle*, et peut être directement utilisée pour l'estimation par maximum de vraisemblance en ignorant la constante cachée qui absorbe les quantités de nuisance et ne modifie en rien la solution. La formulation précédente de la vraisemblance partielle est une conséquence assez naturelle de la structure très spécifique du problème censuré à droite de l'analyse de survie, comme on peut le voir dans figure C.7 ; car elle exprime simplement que lorsque l'observation est la vraie quantité, nous pouvons mettre à jour nos connaissances de la manière habituelle ; en revanche, lorsque l'observation est censurée, le mieux que nous pouvons apprendre est "le vrai temps jusqu'à l'événement est plus grand que l'observation actuelle".

Puisque équation (C.2) rend possible l'estimation par maximum de vraisemblance, il est possible d'aborder le problème de l'analyse de survie par le biais de la modélisation paramétrique, en prenant soin de choisir une famille paramétrique avec support dans $\mathbb{R}_+$ ce qui est par exemple l'approche que nous utiliserons dans chapitre 3. En raison de son lien intime avec la recherche médicale, il est courant en analyse de survie de ne pas s'intéresser à la survie S elle-même, mais à la comparaison de S_1 et S_2, les survies spécifiques à la cause de deux populations, correspondant par exemple à un traitement de référence ou placebo et à un nouveau traitement ou plus généralement de $S(\cdot \mid X_1)$ par rapport à $S(\cdot \mid X_2)$. Dans ce cas, les approches semi-paramétriques ont connu un grand succès dont le modèle des hasards proportionnels de D. R. Cox (1972) est certainement le représentant le plus emblématique. Dans le modèle à risques proportionnels, souvent appelés modèle de Cox, les risques sont supposés être proportionnels de telle sorte que

$$\lambda(t \mid X) = \lambda_0(t) \exp(\theta^\intercal X),$$

où λ_0 est intentionnellement gardé comme non paramétrique et entière-

ment général. La vraisemblance partielle peut alors s'écrire sous la forme

$$\sum_{i:\delta_i=1}\left(\theta^\intercal X_i - \log \sum_{j:T_j\geq T_i} \exp\big(\theta X_j\big)\right),$$

qui, étonnamment, n'impliquent pas λ_0 de quelque manière que ce soit et peux donc être apprise sans difficulté. Notez cependant que si l'on s'intéresse à $\lambda(\cdot \mid X)$ et pas seulement à $\lambda(\cdot \mid X_1)/\lambda(\cdot \mid X_2)$, il est alors possible d'estimer λ_0 de manière non paramétrique (Breslow [1975]).

De même, il existe une abondante littérature sur les modèles de régression tels que le modèle AFT (Buckley et James [1979]) où Y est modélisé comme

$$\log(Y) = -\log\big(f(X)\big) + \epsilon,$$

avec ϵ une distribution de base, ou comme une régression de Poisson.[167] Ces nombreuses méthodes présentent toutefois des défauts importants que nous aimerions éviter. La plupart des résultats de la littérature statistique font l'hypothèse que le modèle estimé et le vrai modèle génératif sont dans la même classe ce qui n'est, bien sûr, jamais vrai, mais a souvent été accepté comme inévitable afin d'obtenir des résultats théoriques intéressants. Nous préférerions cependant des résultats qui correspondent à la réalité des données, c'est-à-dire des bornes qui ne supposent pas que le modèle est correct, même si les bornes résultantes sont nécessairement moins serrées. De même, la plupart des résultats traitent de la convergence, et de sa caractérisation, dans le régime asymptotique. Même si les résultats obtenus ainsi sont très puissants, ils sont peu utiles aux praticiens confrontés à des échantillons de taille finis. Ces deux dernières remarques constituent la base de la théorie de l'*apprentissage statistique*, ou de ce que la plupart des gens ont fini par appeler l'*apprentissage automatique*. Ce travail tente donc de rapprocher le monde de l'analyse de survie et celui de l'apprentissage automatique afin d'apporter aux outils existants de l'analyse de survie le type de garanties théoriques que les praticiens de l'apprentissage automatique en sont venus à attendre.

167 : L'estimateur de Kaplan-Meier du paragraphe précédent peut en fait également être obtenu par une estimation non paramétrique du maximum de vraisemblance

C.3 Prédiction censurée en grande dimension

Comme nous l'avons mentionné dans la section précédente, l'analyse de survie en tant que domaine est principalement née de la nécessité de décrire et de comprendre des phénomènes naturels. Le court interlude historique du domaine donné dans annexe B donne plusieurs exemples d'utilisation de modèles dans le but de mieux comprendre le monde afin de prendre des décisions. Cette approche de la modélisation est certainement la plus naturelle pour la plupart des gens, car elle est à la fois historique

et, surtout, celle à laquelle les gens ont été exposés au cours de leur vie. *Décrire* la nature est, bien sûr, de la plus haute importance pour les épidémiologistes, les virologues, les économétristes ou tout autre domaine scientifique, mais, en général, les praticiens industriels se contentent de résultats beaucoup plus simples, mais pratiques. Lorsqu'on essaie de lancer un ballon de basket dans un panier, il est certainement utile de comprendre la mécanique newtonienne pour savoir que la balle suivra une parabole, mais il est plus que suffisant de prédire que la balle ira simplement *dans cette direction* si vous la lancez *de cette façon* sans rien comprendre aux lois du mouvement. Le même principe s'applique à de nombreux domaines analytiques et, dans notre cas, à la médecine et à la finance. Si l'objectif est simplement de prédire la survenue d'un événement, et non de comprendre les raisons menant à cet événement, il suffit d'adopter un point de vue *prédictif*. Nous appelons ici *prediction* la tâche de deviner, c'est-à-dire de construire un estimateur, une certaine quantité Y à partir de l'entrée ou des caractéristiques X en supposant que $Y = f(X)$. Pour notre basketteur, X est l'angle et la force du lancer alors que dans le cadre médical, X serait les caractéristiques du patient. De ce point de vue, la fonction f est une boîte noire abstraite englobant toute la dynamique menant au résultat, car seul le résultat $Y = f(X)$ nous intéresse. Nous avons précédemment donné l'exemple du modèle de régression de Cox, où la survie d'un individu est modélisée par le taux de hasard instantané λ tel que

$$\lambda(t \mid X) = \lambda_0(t) \exp\left(\beta^\intercal X\right).$$

Si ce modèle peut et est souvent utilisé comme modèle prédictif, sa raison d'être première est l'étude de l'impact relatif des différentes variables à travers l'étude des coefficients β_i ; le but étant de *comprendre* les mécanismes conduisant à la mort. En revanche, si le but est seulement de deviner la quantité $Y = f(X)$, il est suffisant et plus simple[168] de décider d'un certain critère de qualité de la perte $\mathcal{L}$ afin d'essayer de trouver le meilleur f possible pour ce critère compte tenu des données. Mathématiquement, nous pouvons exprimer ce vague objectif en répondant à la question suivante

$$\operatorname*{argmin}_f \mathbb{E}\left[\mathcal{L}\left(Y, f(X)\right)\right], \tag{C.3}$$

qui est souvent appelé le problème de *minimisation du risque*. Cette formulation, bien que peu naturelle à première vue, englobe en réalité de nombreuses questions courantes sur les données que l'on peut avoir en fonction du choix de $\mathcal{L}$. Par exemple, si l'on considère que $\mathcal{L}$ est la perte quadratique $(Y - f(X))^2$, la solution de équation (C.3) est l'espérance conditionnelle $\mathbb{E}[Y \mid X]$, tandis que la valeur absolue $|Y - f(X)|$ conduit à la médiane. De même, des quantités telles que les quantiles ou la probabilité conditionnelle peuvent être obtenues de manière similaire en choisissant

168 : Nous utilisons ici *simple* pour signifier que nous pouvons nous attendre en pratique à de meilleurs résultats sur cette tâche en utilisant les mêmes données, par rapport à l'estimation de la distribution et à la formation d'un régresseur plugin.

169 : Respectivement en choisissant la perte *pinball* et l'entropie croisée.

des pertes appropriées $\mathcal{L}$[169] et l'*art* de choisir la perte correcte pour une tâche particulière attire une attention considérable de la part du monde de la recherche. Nous soulignons que l'*estimation* i.e. l'apprentissage de la densité conditionnelle ou même de l'espérance conditionnelle n'est pas le but poursuivi ici, et qu'il s'agit de *prédiction* par l'apprentissage d'une règle prédictive f avec de bonnes propriétés de généralisation. Bien que les mêmes objets puissent être impliqués dans les deux objectifs, l'objectif en lui-même n'est pas le même comme l'illustre équation (C.4) dans le cas où la fonction prédictive peut être écrite comme une intégrale, ce qui est par exemple le cas pour la moyenne.

$$\underbrace{\int \varphi(y,x)\,\underbrace{p(y \mid x)}_{\text{Objectif}}\,\mathrm{d}y}_{\text{Sous-produit}} \qquad \underbrace{f(x) = \int \varphi(y,x)\,\underbrace{p(y \mid x)}_{\text{Sous-produit}}\,\mathrm{d}y}_{\text{Objectif}}. \tag{C.4}$$

Bien que de nombreux outils utilisés pour la prédiction soient les mêmes que ceux utilisés dans les approches plus traditionnelles, la différence fondamentale dans la question à laquelle on répond justifie des résultats théoriques différents. Comme nous l'avons dit précédemment, nous nous intéressons à la résolution du problème équation (C.3), c'est-à-dire à la recherche de la meilleure f possible en moyenne parmi une famille $\mathcal{F}$ de fonctions potentielles où notre critère dépend de l'objectif final. Bien entendu, nous ne connaissons pas la distribution de (Y, X) ni même la véritable famille de fonctions qui contient le véritable f, et ne pouvons donc pas résoudre directement équation (C.3). Nous pouvons toutefois résoudre la version empirique de ce problème à partir des données dont nous disposons, à savoir

$$\operatorname*{argmin}_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}\left(Y_i, f(X_i)\right), \tag{C.5}$$

que nous appelons l'approche de *minimisation du risque empirique* (ERM). Étant donné que l'on ne résout pas le bon problème, mais seulement une version empirique et restreinte de celui-ci, il semble légitime de se demander quelles sont les garanties que la solution obtenue soit une bonne solution. Cette dernière question est le principal problème de la *théorie de l'apprentissage statistique* et a été abordée sous de nombreux angles, nous adoptons cependant ici l'approche PAC : nous disons que la solution $\hat{f}_n$ de équation (C.5) est bonne si elle est bonne selon équation (C.5) avec une grande probabilité. C'est-à-dire, étant donné que nous savons seulement calculer

$$\mathcal{R}_n(\hat{f}_n) = \min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}\left(Y_i, f(X_i)\right),$$

que la quantité

$$\left|\mathcal{R}_n(\hat{f}_n) - \mathcal{R}(f^*)\right|, \tag{C.6}$$

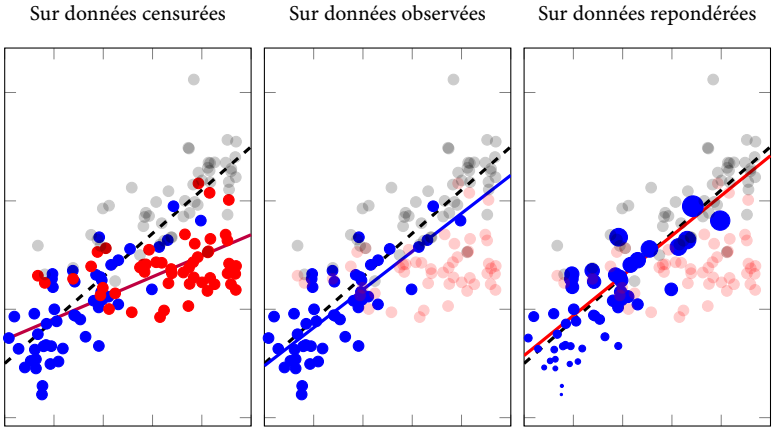


FIGURE C.8 : Apprentissage d'une fonction linéaire sur les données censurées brutes, les données entièrement observées et les données observées repondérées.

est petite,[170] où $\mathcal{R}(f^*)$ est le minimum de équation (C.3). De nombreux résultats de cette forme existent, comme nous le verrons dans, §2.1, mais il reste un problème flagrant qui rend l'approche ERM inadaptée à l'analyse de survie : dans équation (C.5) Y_i n'est pas observé. Il est toutefois possible d'adapter équation (C.5) au cadre de la survie et de prouver des résultats similaires à ceux qui existent déjà dans la théorie de l'apprentissage statistique sans censure.

170 : De même que $\mathcal{R}_n(\hat{f}_n)$ est petite, mais ceci est implicite étant donné qu'elle est explicitement définie telle quelle.

C.3.1 Prédiction censurée

Dans notre cas, Y est inobservée et seuls (T, δ), la variable censurée ainsi que l'indicateur de censure, sont observés à la place. Nous montrons, en suivant la série séminale de Stute (1996, 1993a,b, 1995a,b, 2003) ; Stute et J.-L. Wang (1993) et les travaux de Dabrowska (1989) que la quantité inobservable de équation (C.3) peut être remplacée par la quantité repondérée, mais mathématiquement équivalente suivante :

$$\operatorname*{argmin}_f \mathbb{E}\left[\frac{\delta}{S_C(T- \mid X)}\mathcal{L}\left(T, f(X)\right)\right], \tag{C.7}$$

et la version empirique correspondante

$$\operatorname*{argmin}_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i}{S_C(T_i \mid X_i)}\mathcal{L}\left(T_i, f(X_i)\right). \tag{C.8}$$

Bien sûr, alors que la variable inobservable Y soit remplacée par les quantités observables T et δ, nous faisons maintenant intervenir la fonction de survie S_C qui elle est inconnue. En résolvant ce nouveau problème

empirique repondéré, au lieu de s'appuyer sur l'ensemble des données censurées ou uniquement sur les individus entièrement observés, nous sommes en mesure d'éliminer le biais d'estimation qui résulterait autrement en une sous-estimation, comme on peut le voir dans figure C.8. Bien que ce changement semble inutile puisque nous avons échangé une quantité inconnue contre une autre, nous savons comment estimer S_C comme vu dans annexe C.2 et nous pouvons donc plutôt étudier

$$\operatorname*{argmin}_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \frac{\delta_i}{\hat{S}_C(T_i \mid X_i)} \mathcal{L}\left(T_i, f(X_i)\right). \tag{C.9}$$

Dans le chapitre 2, nous montrons qu'en utilisant un estimateur à noyau $\hat{S}_C$ de S_C, nous pouvons obtenir des bornes non asymptotiques et non paramétriques de l'erreur de généralisation équation (C.6) similaires à celles présentes de la littérature de l'apprentissage statistique sans censure. Comme $\hat{S}_C(T_i \mid X_i)$ est elle-même une variable aléatoire impliquant tous les couples (T_i, X_i) des données d'apprentissage, ici sous la forme d'une somme d'estimateurs indépendants, le rapport

$$\frac{\delta_i}{S_C(T_i \mid X_i)} \mathcal{L}\left(T_i, f(X_i)\right),$$

n'est pas indépendant et identiquement distribué, ce qui rend invalides la plupart des techniques de preuve impliquant des sommes empiriques de variables i.i.d. A défaut, nous nous appuyons sur le fait que la quantité précédente peut être écrite comme un rapport de sommes afin de la linéariser et de la traiter ensuite comme une U-statistique, c.a.d.. une généralisation de la moyenne empirique, sur laquelle des résultats de concentration peuvent être appliqués. Cela nous permet de prouver dans théorème 2.9 des bornes de généralisation sur le problème ERM censuré qui sont similaires au cas complètement observé :

Theorem (Contrôle uniforme de l'excès de risque)**.** *Supposons que les Assumptions 2.1 et 2.4 soient remplies. Il existe des constantes h_0, M_1, M_2 et M_3 qui dépendent uniquement de (A, v), M_Φ, L, K et b, de sorte que, pour tout $n \geq 2$ et $\varepsilon \in (0,1)$, l'événement*

$$\left|\mathcal{R}(\tilde{f}_n) - \mathcal{R}(f^\star)\right| \leq M_1 \left(\sqrt{\frac{\log\left(M_2/\varepsilon\right)}{n}} + \frac{\left|\log(\varepsilon h^{d/2})\right|}{nh^d} + h^2 \right),$$

se produit avec une probabilité supérieure à $1 - \varepsilon$ à condition que $h \leq h_0$, $nh^{2d} \geq M_3 \left|\log(\varepsilon h^d)\right|$.

De plus, nous prouvons expérimentalement dans §2.5 que les performances obtenues sur données réelles avec le cadre proposé correspondent à celles attendues à partir des bornes théoriques.

Ces résultats, qui représentent la principale contribution de cette thèse, ont été présentés de manière préliminaire lors de l'atelier Machine Learning for Health à NeurIPS 2018 (Ausset, Portier et Clémençon [2018]) et sont en cours de révision finale pour publication au JMLR au moment de la rédaction de cet article.

Papiers de chapitre 2

Guillaume Ausset, François Portier et Stéphan Clémençon (2018). « Machine Learning for Survival Analysis : Empirical Risk Minimization for Censored Distribution Free Regression with Applications ». In : NeurIPS ML4H Workshop. Montreal, Canada

```
@inproceedings{aussetMachineLearningSurvival2018,
  title = {Machine {{Learning}} for {{Survival Analysis}}:
    {{Empirical Risk Minimization}} for
    {{Censored Distribution Free Regression}} with {{Applications}}},
  author = {Ausset, Guillaume and Portier, François and Clémençon, Stéphan},
  date = {2018},
  location = {{Montreal, Canada}},
  url = {https://hal.archives-ouvertes.fr/hal-02287991},
  eventtitle = {{{NeurIPS ML4H Workshop}}}
}
```

Guillaume Ausset, Stéphan Clémençon et François Portier (2021a). « Empirical Risk Minimization under Random Censorship ». *under revision in Journal of Machine Learning Research*. arXiv : 1906.01908

```
@article{aussetEmpiricalRiskMinimization2021,
  title = {Empirical {{Risk Minimization}} under {{Random Censoring}}},
  shorttitle = {Empirical {{Risk Minimization}} under {{Random Censoring}}},
  author = {Ausset, Guillaume and Clémençon, Stéphan and Portier, François},
  date = {2021},
  journaltitle = {Journal of Machine Learning Research (Provisional)},
  url = {http://arxiv.org/abs/1906.01908}
}
```

C.3.2 Estimateurs flexibles de la survie

Bien que les résultats présentés dans le chapitre 2 donnent de solides justifications théoriques pour l'utilisation du cadre IPCW ERM, la performance dépend toujours fortement de la qualité des poids $\delta_i/S_C(T_i|X_i)$ et donc de l'estimateur de S_C. Au-delà de l'utilisation dans la régression IPCW, les estimateurs de la survie présentent pour la communauté un intérêt en soi, et de nombreux estimateurs flexibles ont été proposés au fil des ans. Dans le chapitre 3, basé sur Ausset, Ciffreo et al. (2021), nous étudions un type particulier d'estimateur de S construit à partir d'un modèle génératif de la variable d'intérêt, c.a.d.. Y dans le cadre général de la survie ou C pour les

poids IPCW, avec une vraisemblance accessible. En modélisant Y comme la variable transformée

$$Y = m_\theta(Z, X), \tag{C.10}$$

où m_θ est une famille flexible de réseaux neuronaux paramétrés par $\theta \in \Theta$ et Z est une distribution simple connue. Nous sommes capables de déterminer le m_θ optimal en maximisant la log-vraisemblance censurée

$$\sum_{i=1}^{n} \Big(\delta_i p_{Y,\theta}\left(T_i \mid X_i\right) + (1 - \delta_i) S_{Y,\theta}\left(T_i \mid X_i\right) \Big), \tag{C.11}$$

où $p_{Y,\theta}$ et $S_{Y,\theta}$ sont la densité et la fonction de survie de Y telles que paramétrées par équation (C.10). La paramétrisation donnée par équation (C.10) est un type de modèle génératif introduit pour la première fois sous le nom de *flux normalisant* par Rezende et Mohamed (2015). Son utilité réside dans le fait que $p_{Y,\theta}$ peut être obtenu à partir de p_Z au moyen de la formule de changement de variable

$$\log p_{Y,\theta}(t \mid X) = \log p_Z(z) - \log \left| \det \frac{\partial m_\theta}{\partial z} \right|.$$

De même, nous montrons dans le chapitre 3 qu'il est également possible de retrouver la survie $S_{Y,\theta}$ en adoptant la formulation continue de équation (C.12) (voir R. T. Q. Chen et al. [2018]),

$$\begin{aligned} \frac{\partial}{\partial t} \begin{bmatrix} \mathbf{z}_\theta(t, X) \\ \log p(y \mid X) - \log p(\mathbf{z}_\theta(t, X)) \end{bmatrix} &= \begin{bmatrix} m_\theta(\mathbf{z}_\theta(t, X), t, X) \\ -\operatorname{tr} \dfrac{\partial m_\theta}{\partial \mathbf{z}} \end{bmatrix}, \\ \begin{bmatrix} \mathbf{z}_\theta(1, X) \\ \log p(y \mid X) - \log p(\mathbf{z}_\theta(1, X)) \end{bmatrix} &= \begin{bmatrix} y \\ 0 \end{bmatrix}, \end{aligned} \tag{C.12}$$

permettant de calculer ainsi que de dériver (voir Rackauckas, Ma, Dixit et al. [2018], pour la différentiabilité des solutions d´ODEs) toutes les quantités présentes dans équation (C.11).

Malgré le coût de calcul élevé de la méthode proposée, nous montrons que par rapport aux approches neuronales existantes telles que DeepCox (Nagpal et al. [2021]) ou DeepHit (C. Lee, Zame et al. [2018] ; C. Lee, Yoon et Schaar [2020]), cette approche CNF présente des performances compétitives sur les tâches de régression classiques, mais autorise également de nouvelles applications. En tant que modèle génératif, l'approche CNF permet de tirer efficacement des observations conditionnelles, une caractéristique très utile en finance où les stress tests et les simulations sont des exigences réglementaires, mais aussi pour des applications où des dépendances complexes doivent être modélisées et simulées par Monte-Carlo. Cette dernière application, étant donné son importance particulière pour la finance, est étudiée plus en détail dans le chapitre 5.

Si les avantages des modèles génératifs de survie flexibles basés sur les réseaux de neurones sont indéniables, le coût de calcul peut être difficile à justifier si l'on considère les performances relativement élevées de méthodes plus simples, et presque gratuites en comparaison, telles que les forêts aléatoires de survie (Ishwaran et Kogalur [2007]) ou même le modèle de Cox (D. R. Cox et Oakes [1984]) ; même si l'on peut contraster par le fait que la majeure partie du coût est encourue pendant l'apprentissage et amortie pendant l'inférence. Malgré le fait que la méthode CNF soit intrinsèquement plus coûteuse, il est toujours possible d'atténuer la charge de calcul en réduisant la taille du réseau neuronal, ce palliatif ne pouvant fonctionner que si le nombre de dimensions de X lui-même est réduit en même temps. Par conséquent, pour que la méthode proposée présente un intérêt pratique, il faut trouver un moyen robuste de réduire la dimension de X.

Papiers de chapitre 3

Guillaume Ausset, Tom Ciffreo et al. (2021). « Individual Survival Curves with Conditional Normalizing Flows ». In : *DSAA'21*. IEEE International Conference on Data Science and Advanced Analytics

```
@inproceedings{aussetIndividualSurvivalCurves2021,
  title = {Individual {{Survival Curves}}
    with {{Conditional Normalizing Flows}}},
  booktitle = {{{DSAA}}'21},
  author = {Ausset, Guillaume and Ciffreo, Tom
    and Clémençon, Stéphan and Portier, François and Papin, Timothée},
  date = {2021},
  eventtitle = {{{IEEE International Conference}} on {{Data Science}}
    and {{Advanced Analytics}}}
}
```

C.3.3 Gestion de la grande dimension

La dernière contribution majeure de cette thèse présentée dans le chapitre 4 est une réponse au besoin précédemment mentionné d'une technique robuste de réduction de la dimension. Le but de la réduction de la dimension est de trouver un espace de dimension inférieure qui capture la majorité de l'information présente dans l'espace d'origine. La notion même d'*information* est laissée ici intentionnellement assez vague, car, selon la tâche ou les besoins spécifiques du problème, elle peut changer radicalement conduisant ainsi à des techniques très différentes de réduction de la dimension.

Nous pouvons, très rudimentairement, diviser les types de tâches en deux groupes distincts [171] : *non supervisé* et *supervisé*. Par *non super-*

171 : Mais avec des recoupements importants.

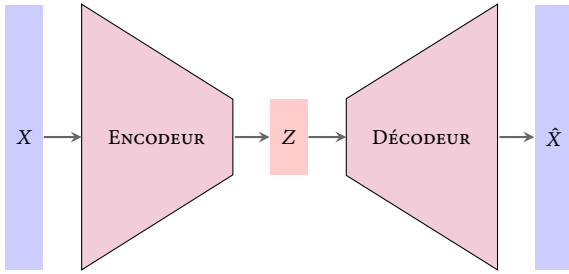


Figure C.9 : VAE simplifié.

visé, nous entendons les tâches où l'objet d'intérêt est X lui-même qui est considéré comme la seule quantité observée.[172] Sans aucune information auxiliaire, la façon la plus naturelle de formuler le problème est donc de le considérer simplement comme un problème de reconstruction, qui consiste à trouver un espace de dimension inférieure obéissant à certaines contraintes supplémentaires, de sorte que ce nouvel espace minimise une certaine notion de distance par rapport à l'espace d'origine. C'est par exemple l'approche adoptée pour l'analyse en composante principale (ACP) où la perte quadratique d'une projection sur un sous-espace de dimension l est minimisée.[173] De même les VAEs trouvent la représentation latente Z de dimension inférieure l en apprenant conjointement la fonction de plongement, ou encodeur, enc et la fonction de décodage dec qui minimisent l'erreur de reconstruction

$$\|X - \text{dec} \circ \text{enc}(X)\|,$$

ce qui peut être considéré comme une généralisation de l'ACP, en prenant enc = P une projection et dec = id,[174] comme illustré dans figure C.9.[175] D'autre part, les techniques de réduction de la dimension *supervisées* s'attachent à trouver de bonnes représentations de X lorsque (X, Y) est observé et que la prédiction de Y étant donné X est la tâche d'intérêt. Il est par exemple possible de considérer l'analyse discriminante linéaire (ADL) comme une extension supervisée de l'ACP[176] qui, au lieu de trouver la projection qui maximise la variance, trouve la projection qui maximise la séparation des classes. De même, par analogie avec l'approche VAE, au lieu de trouver une représentation adaptée à la reconstruction, il est habituel dans le cadre supervisé de trouver une représentation adaptée à la prédiction, ou à la classification, en prenant simplement l'avant-dernière couche d'un réseau de neurones, avant la couche de sortie EC, comme représentation de dimension inférieure d'intérêt tel que représenté dans figure C.10.

Une autre approche du problème supervisé consiste à considérer comme importantes les variables qui ont un impact sur la sortie

$$Y = f(X) + \varepsilon,$$

172 : Même dans ce cas, certaines personnes considèrent cette tâche comme un problème *autosupervisé*, c'est-à-dire où les covariables sont X et la variable dépendante également X.

173 : Ceci s'avère etre équivalent à la maximisation de la variance sur le sous-espace projeté.

174 : Ceci n'est pas une revue exhaustive, plus de détails sont donnés dans le chapitre 4. Notez que si toutes ces techniques peuvent être considérées d'un point de vue de reconstruction pure, c'est-à-dire comme des problèmes d'optimisation, elles admettent également des interprétations probabilistes.

175 : La formulation réelle des VAEs est probabiliste et variationnelle par essence, car la formulation donnée ici surapprendra les données.

176 : La ADL est souvent considérée comme un algorithme de classification, mais elle peut être utilisée pour la réduction de dimension.

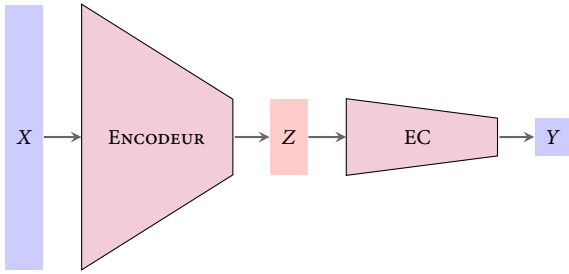


Figure C.10 : Avant-dernière couche d'un réseau de neurones comme représentation.

ce qui, assez naïvement, peut être considéré comme la recherche des dérivées directionnelles non nulles. Si l'on se limite uniquement à la sélection des variables, c'est-à-dire en restreignant les directions possibles aux axes seuls, alors le problème est simplement de trouver les éléments non nuls du gradient. Cette approche a été étudiée et justifiée dans la littérature dans le cadre du *simple* et *multi-index* où

$$Y = f(TX) + \varepsilon,$$

avec T une matrice de projection. Dans ce modèle, il est clair que le sous-espace EDR défini par T est engendré par le gradient ∇r de $r(x) = f(Tx)$. Plusieurs approches à ce problème spécifique ont déjà été proposées dans la littérature et sont décrites en détail dans, §4.1, mais aucune ne combine des bornes non asymptotiques et uniformes sur l'erreur du gradient et de la fonction de régression elle-même avec une approche des k plus proches voisins (k-NN) lorsque le gradient est supposé creux. L'approche k-NN est particulièrement intéressante en pratique, car elle est non seulement facile à calibrer et à comprendre pour les profanes, mais elle permet également d'éviter les cas pathologiques où trop peu d'exemples sont présents dans un voisinage choisi. Nous donnons dans le chapitre 4 une formulation linéaire locale LASSO du problème de la forme

$$\underset{(r,\beta)\in\mathbb{R}^{d+1}}{\operatorname{argmin}} \sum_{i\in\hat{i}_k(x)} \left(Y_i - r - \beta^\intercal (X_i - x)\right)^2 + \lambda\|\beta\|_1, \tag{C.13}$$

et montrons dans théorème 4.1 qu'il est possible d'exploiter la sparsité supposée du gradient pour améliorer les bornes de l'erreur, avec des résultats similaires sur la fonction de régression elle-même dans théorème 4.2.

Theorem. *Supposons que Assumptions 4.1 à 4.4 soient remplies. Soient $n \geq 1$ et $k \geq 1$ tels que $C_k \leq C_0$, prenons*

$$\lambda = C_k \left(\sqrt{2\sigma^2 \frac{\log(16d/\delta)}{k}} + LC_k^2 \right).$$

Alors, nous avons avec une probabilité plus grande que $1-\delta$,

$$\left\|\tilde{\nabla}_k r(x) - \nabla r(x)\right\|_2 \leq 24^2 \sqrt{|S_x|}\left(C_k^{-1}\sqrt{\frac{2\sigma^2 \log(16d/\delta)}{k}} + LC_k\right),$$

dès que

$$C_1 |S_x| \log\left(\frac{dn}{\delta}\right) \leq k \leq C_2 n,$$

où $\tilde{\nabla}_k r(x)$ *est la deuxième composante de la solution de équation* (C.13), C_0, C_1, C_2 *et* L *sont des constantes universelles,* C_k *est une constante définie en détail dans théorème 4.1 et* $|S_x|$ *est le nombre de composantes non nulles de* $\nabla r(x)$.

La plupart des exemples de réduction de dimension donnés précédemment, telles que ACP ou ALD, supposent que les variables importantes sont les mêmes pour tous les individus et sont donc traitées comme une étape de prétraitement appliquée à l'ensemble des données avant toute autre analyse. Il n'y a cependant aucune raison pour qu'une hypothèse aussi forte soit vraie : si l'ensemble de données comprend par exemple des hommes et des femmes, il semble pour le moins douteux de penser que les mêmes variables sont importantes pour les deux sexes et on s'attendrait à ce qu'une méthode de sélection de variables adaptative soit plus performante. De même, lorsqu'on compare les caractéristiques des clients d'un prêt, il semble souhaitable de prendre en compte des caractéristiques différentes si le client est une multinationale ou une petite entreprise locale. Comme notre méthode est locale, c'est-à-dire que le gradient est estimé à un x spécifique et récupère donc les variables d'importance dans un voisinage de x, il est possible de sélectionner différentes variables pertinentes dans différentes régions de l'espace. Dans §4.5, nous montrons non seulement comment trouver des variables *globalement* importantes en agrégeant les gradients en toutes les observations, mais nous proposons également une méthode à base d'arbres exploitant l'information locale du gradient afin d'améliorer les performances.

Enfin, bien que l'analyse théorique du chapitre soit effectuée en ignorant momentanément toute forme de censure pour des raisons de simplicité, il s'agit toujours d'un problème ERM sur lequel on peut appliquer l'approche IPCW du chapitre 2, par exemple pour identifier les gènes responsables de la survie d'un cancer comme fait dans §4.5.2.

Papiers de chapitre 4

Guillaume Ausset, Stéphan Clémençon et François Portier (2021b). « Nearest Neighbour Based Estimates of Gradients : Sharp Nonasymptotic Bounds and Applications ». In : *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics*. Sous la dir. d'Arindam Banerjee et Kenji Fukumizu. T. 130. Proceedings of Machine Learning Research. PMLR, p. 532-540

```
@inproceedings{aussetNearestNeighbourBased2021,
  title = {Nearest Neighbour Based Estimates of Gradients:
    {{Sharp}} Nonasymptotic Bounds and Applications},
  booktitle = {Proceedings of the 24th International Conference
    on Artificial Intelligence and Statistics},
  author = {Ausset, Guillaume and Clémencon, Stephan and Portier, François},
  date = {2021},
  series = {Proceedings of Machine Learning Research},
  volume = {130},
  pages = {532--540},
  publisher = {{PMLR}},
  url = {http://proceedings.mlr.press/v130/ausset21a.html}
}
```

C.4 Schéma de ce manuscrit

Ce manuscrit est organisé comme suit :

- Chapitre 2 traite du cadre de la minimisation du risque empirique en présence de censure. Des bornes supérieures non asymptotiques et uniformes de l'erreur de généralisation sont prouvées, tandis que la fin du chapitre est consacrée aux expériences numériques justifiant la validité de l'approche au-delà des résultats théoriques.
- Chapitre 3 présente les flux normalisant pour l'analyse de survie, un modèle génératif dont la vraisemblance est accessible. Plusieurs applications servent d'exemple de l'utilité d'une telle approche dans le cadre classique de la survie, tandis que la motivation de la formulation générative est étudiée plus en détail dans chapitre 5.
- Chapitre 4 traite du problème de la grande dimension à travers la sélection de variables. Un estimateur du gradient est introduit et des bornes non asymptotiques de l'erreur du gradient supposé creux ainsi que du régresseur sont prouvées. Comme le gradient est lui-même utile au-delà de la simple sélection de variables, plusieurs exemples d'optimisation d'ordre zéro ainsi que de désenchevêtrement sont donnés.
- Chapitre 5 traite de l'analyse de survie en finance à travers le cas spécifique de la titrisation, l'une des activités principales de BNP

Paribas CIB. Nous motivons le modèle génératif introduit dans chapitre 3 par l'utilisation de la modélisation hiérarchique multi-niveaux.

Au-delà du texte principal qui introduit les contributions de cette thèse, quelques détails supplémentaires sont donnés en annexe.

- Les preuves classiques qui ne sont pas strictement nécessaires à la compréhension des résultats principaux sont données dans annexe A pour référence.
- Un aperçu historique de l'analyse de survie est donné en guise de distraction et d'intermède aux preuves techniques dans annexe B.
- Chapitre 1 contient la version anglaise originale de cette introduction.

Bibliography

Aalen, Odd (1978). "Nonparametric Inference for a Family of Counting Processes". *Annals of Statistics* 6:4, pp. 701–726. DOI: 10.1214/aos/1176344247 (cit. on pp. 9, 110, 214).

Abolfathi, Bela et al. (19, 2018). "The Fourteenth Data Release of the Sloan Digital Sky Survey: First Spectroscopic Data from the Extended Baryon Oscillation Spectroscopic Survey and from the Second Phase of the Apache Point Observatory Galactic Evolution Experiment". *The Astrophysical Journal Supplement Series* 235:2, p. 42. DOI: 10.3847/1538-4365/aa9e8a (cit. on p. 150).

Agrawal, Akshay, Brandon Amos, Shane Barratt, Stephen Boyd, Steven Diamond, and J. Zico Kolter (2019). "Differentiable Convex Optimization Layers". In: *Advances in Neural Information Processing Systems 32*. Ed. by H. Wallach, H. Larochelle, A. Beygelzimer, Florence d'Alché-Buc, E. Fox, and R. Garnett. Curran Associates, Inc., pp. 9562–9574 (cit. on pp. 120, 158).

Alman, Josh and Virginia Vassilevska Williams (12, 2020). *A Refined Laser Method and Faster Matrix Multiplication*. arXiv: 2010.05846 [cs, math]. URL: http://arxiv.org/abs/2010.05846 (cit. on p. 120).

Andersen, Per Kragh, Ørnulf Borgan, Richard D. Gill, and Niels Keiding (1993). *Statistical Models Based on Counting Processes*. Springer Series in Statistics. Springer US, New York, NY. DOI: 10.1007/978-1-4612-4348-9 (cit. on pp. 34, 109).

Artzner, Philippe, Freddy Delbaen, Jean-Marc Eber, and David Heath (1999). "Coherent Measures of Risk". *Mathematical Finance* 9:3, pp. 203–228. DOI: 10.1111/1467-9965.00068 (cit. on p. 179).

Aswani, Anil, Peter Bickel, and Claire Tomlin (2011). "Regression on Manifolds: Estimation of the Exterior Derivative". *The Annals of Statistics* 39:1, pp. 48–81. DOI: 10.1214/10-AOS823. arXiv: 1103.1457 (cit. on p. 160).

Aumüller, Martin, Erik Bernhardsson, and Alexander Faithfull (17, 2018). *ANN-Benchmarks: A Benchmarking Tool for Approximate Nearest Neighbor Algorithms*. arXiv: 1807.05614 [cs]. URL: http://arxiv.org/abs/1807.05614 (cit. on p. 148).

Ausset, Guillaume, Tom Ciffreo, Stéphan Clémençon, François Portier, and Timothée Papin (2021). "Individual Survival Curves with Conditional Normalizing Flows". In: *DSAA'21*. IEEE International Conference on Data Science and Advanced Analytics (cit. on pp. 16, 18, 107, 108, 221, 223).

Ausset, Guillaume, Stéphan Clémençon, and François Portier (2021a). "Empirical Risk Minimization under Random Censorship". *under revision in Journal of Machine Learning Research*. arXiv: 1906.01908 (cit. on pp. 16, 33, 221).

Ausset, Guillaume, Stéphan Clémençon, and François Portier (2021b). "Nearest Neighbour Based Estimates of Gradients: Sharp Nonasymptotic Bounds and Applications". In: *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics*. Ed. by Arindam Banerjee and Kenji Fukumizu. Vol. 130. Proceedings of Machine Learning Research. PMLR, pp. 532–540 (cit. on pp. 21, 138, 227).

Ausset, Guillaume, François Portier, and Stéphan Clémençon (2018). "Machine Learning for Survival Analysis: Empirical Risk Minimization for Censored Distribution Free Regression with Applications". In: NeurIPS ML4H Workshop. Montreal, Canada (cit. on pp. 15, 16, 221).

Avati, Anand, Tony Duan, Sharon Zhou, Kenneth Jung, Nigam H. Shah, and Andrew Ng (21, 2018). *Countdown Regression: Sharp and Calibrated Survival Predictions*. arXiv: 1806.08324 [cs, stat]. URL: http://arxiv.org/abs/1806.08324 (cit. on p. 118).

Bach, Francis R. and Michael I. Jordan (2005). "Predictive Low-Rank Decomposition for Kernel Methods". In: *Proceedings of the 22nd International Conference on Machine Learning - ICML '05*. The 22nd International Conference. ACM Press, Bonn, Germany, pp. 33–40. DOI: 10.1145/1102351.1102356 (cit. on p. 149).

Bang, Heejung and Anastasios A. Tsiatis (2002). "Median Regression with Censored Cost Data". *Biometrics* 58:3, pp. 643–649. JSTOR: 3068588 (cit. on p. 39).

Bareiss, Erwin H. (1968). "Sylvester's Identity and Multistep Integer-Preserving Gaussian Elimination". *Mathematics of Computation* 22:103, pp. 565–578. DOI: 10.2307/2004533. JSTOR: 2004533 (cit. on p. 120).

Bartlett, Peter L., Olivier Bousquet, and Shahar Mendelson (2005). "Local Rademacher Complexities". *Annals of Statistics* 33:4, pp. 1497–1537. DOI: 10.1214/009053605000000282 (cit. on p. 32).

Basel Committee (17, 2018). *Stress Testing Principles*. Basel: Bank for International Settlements (cit. on p. 117).

– (2019). *Scope and Definitions*, p. 41 (cit. on pp. 4, 208).

Baudat, G. and F. Anouar (2000). *Generalized Discriminant Analysis Using a Kernel Approach* (cit. on p. 137).

Bellman, Richard (1, 1954). "The Theory of Dynamic Programming". *Bulletin of the American Mathematical Society* 60:6, pp. 503–516. DOI: 10.1090/S0002-9904-1954-09848-8 (cit. on p. 134).

Bengio, Yoshua, Aaron Courville, and Pascal Vincent (23, 2014). *Representation Learning: A Review and New Perspectives*. arXiv: 1206.5538 [cs]. URL: http://arxiv.org/abs/1206.5538 (cit. on p. 136).

Berahas, Albert S., Liyuan Cao, Krzysztof Choromanski, and Katya Scheinberg (1, 2020). *A Theoretical and Empirical Comparison of Gradient Approximations in Derivative-Free Optimization*. arXiv: 1905.01332 [math]. URL: http://arxiv.org/abs/1905.01332 (cit. on pp. 140, 152).

Beran, Rudolf (1981). *Nonparametric Regression with Randomly Censored Survival Data* (cit. on pp. 37, 39).

Bernoulli, Daniel (1766). "Essai d'une Nouvelle Analyse de La Mortalité Causée Par La Petite Vérole et Des Avantages de l'inoculation Pour La Prévenir." *Mémoires de Mathématique et de Physique, présentés à l'Académie Royale des Sciences* (cit. on p. 201).

Bezanson, Jeff, Alan Edelman, Stefan Karpinski, and Viral B. Shah (1, 2017). "Julia: A Fresh Approach to Numerical Computing". *SIAM Review* 59:1, pp. 65–98. DOI: 10.1137/141000671 (cit. on p. 128).

Biau, Gérard, Frédéric Cérou, and Arnaud Guyader (1, 2010). "Rates of Convergence of the Functional K-Nearest Neighbor Estimate". *IEEE Transactions on Information Theory* 56:4, pp. 2034–2040. DOI: 10.1109/TIT.2010.2040857 (cit. on p. 141).

Biau, Gérard and Luc Devroye (2015). *Lectures on the Nearest Neighbor Method*. Springer Series in the Data Sciences. Springer International Publishing. DOI: 10.1007/978-3-319-25388-6 (cit. on p. 139).

Binkowski, Mikolaj, Gautier Marti, and Philippe Donnat (3, 2018). "Autoregressive Convolutional Neural Networks for Asynchronous Time Series". In: *Proceedings of the 35th International Conference on Machine Learning*. International Conference on Machine Learning. PMLR, pp. 580–589 (cit. on p. 117).

Bishop, Christopher M. (2006). *Pattern Recognition and Machine Learning*. DOI: 10.1641/B580519. pmid: 18292226 (cit. on p. 32).

Blondel, Mathieu, Olivier Teboul, Quentin Berthet, and Josip Djolonga (29, 2020). *Fast Differentiable Sorting and Ranking*. arXiv: 2002.08871 [cs, stat]. URL: http://arxiv.org/abs/2002.08871 (cit. on pp. 62, 120).

Bogacki, P. and L. Shampine (1989). "A 3(2) Pair of Runge - Kutta Formulas". DOI: 10.1016/0893-9659(89)90079-7 (cit. on p. 131).

Boucheron, Stéphane, Gábor Lugosi, and Pascal Massart (2013). *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press (cit. on pp. 32, 160, 161).

Bousdekis, Alexandros, Katerina Lepenioti, Dimitris Apostolou, and Gregoris Mentzas (1, 2019). "Decision Making in Predictive Maintenance: Literature Review and Research Agenda for Industry 4.0". *IFAC-PapersOnLine*. 9th IFAC Conference on Manufacturing Modelling, Management and Control MIM 2019 52:13, pp. 607–612. DOI: 10.1016/j.ifacol.2019.11.226 (cit. on pp. 2, 206).

Brault, Romain, Alex Lambert, Zoltan Szabo, Maxime Sangnier, and Florence d'Alche-Buc (11, 2019). "Infinite Task Learning in RKHSs". In: *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*. The 22nd International Conference on Artificial Intelligence and Statistics. PMLR, pp. 1294–1302 (cit. on p. 188).

Breiman, Leo (1, 2001). "Random Forests". *Machine Learning* 45:1, pp. 5–32. DOI: 10.1023/A:1010933404324 (cit. on p. 142).

Breiman, Leo, Jerome H. Friedman, Richard A. Olshen, and Charles J. Stone (1984). *Classification And Regression Trees*. Routledge, Boca Raton. 368 pp. DOI: 10.1201/9781315139470 (cit. on p. 150).

Breslow, N. E. (1975). "Analysis of Survival Data under the Proportional Hazards Model". *International Statistical Review / Revue Internationale de Statistique* 43:1, pp. 45–57. DOI: 10.2307/1402659. JSTOR: 1402659 (cit. on pp. 11, 111, 216).

Brier, Glenn W. and Roger A. Allen (1951). "Verification of Weather Forecasts". In: *Compendium of Meteorology: Prepared under the Direction of the Committee on the Compendium*

of Meteorology. Ed. by H. R. Byers et al. American Meteorological Society, Boston, MA, pp. 841–848. DOI: 10.1007/978-1-940033-70-9_68 (cit. on p. 62).

Brock, Andrew, Jeff Donahue, and Karen Simonyan (25, 2019). *Large Scale GAN Training for High Fidelity Natural Image Synthesis*. arXiv: 1809.11096 [cs, stat]. URL: http://arxiv.org/abs/1809.11096 (cit. on p. 115).

Buckley, Jonathan and Ian James (1979). "Linear Regression with Censored Data". *Biometrika* 66:3, pp. 429–436. DOI: 10.2307/2335161. JSTOR: 2335161 (cit. on pp. 11, 50, 123, 216).

Cao, Yang, Shengtai Li, Linda Petzold, and Radu Serban (1, 2003). "Adjoint Sensitivity Analysis for Differential-Algebraic Equations: The Adjoint DAE System and Its Numerical Solution". *SIAM Journal on Scientific Computing* 24:3, pp. 1076–1089. DOI: 10.1137/S1064827501380630 (cit. on p. 125).

Chen, Yi-Cheng, Ping-En Lu, Cheng-Shang Chang, and Tzu-Hsuan Liu (2020). "A Time-Dependent SIR Model for COVID-19 With Undetectable Infected Persons". *IEEE Transactions on Network Science and Engineering* 7:4, pp. 3279–3294. DOI: 10.1109/TNSE.2020.3024723 (cit. on p. 203).

Chen, Chong et al. (1, 2020). "Predictive Maintenance Using Cox Proportional Hazard Deep Learning". *Advanced Engineering Informatics* 44, p. 101054. DOI: 10.1016/j.aei.2020.101054 (cit. on pp. 2, 206).

Chen, Ricky T. Q., Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud (2018). "Neural Ordinary Differential Equations". In: *Advances in Neural Information Processing Systems*. Vol. 31. Curran Associates, Inc. (cit. on pp. 17, 107, 121, 126, 222).

Clémençon, Stéphan, Gábor Lugosi, and Nicolas Vayatis (2008). "Ranking and Empirical Minimization of U-Statistics". *Annals of Statistics* 36:2, pp. 844–874. DOI: 10.1214/009052607000000910 (cit. on p. 48).

Clémençon, Stéphan and François Portier (31, 2018). "Beating Monte Carlo Integration: A Nonasymptotic Study of Kernel Smoothing Methods". In: *International Conference on Artificial Intelligence and Statistics*. International Conference on Artificial Intelligence and Statistics. PMLR, pp. 548–556 (cit. on p. 39).

Colombo, Camilla and Mirko Diamanti (1, 2015). "The Smallpox Vaccine: The Dispute between Bernoulli and d'Alembert and the Calculus of Probabilities". *Lettera Matematica* 2:4, pp. 185–192. DOI: 10.1007/s40329-015-0073-5 (cit. on pp. 200, 201).

Cooper, Ian, Argha Mondal, and Chris G. Antonopoulos (2020). "A SIR Model Assumption for the Spread of COVID-19 in Different Communities". *Chaos, Solitons, and Fractals* 139, p. 110057. DOI: 10.1016/j.chaos.2020.110057. pmid: 32834610 (cit. on p. 203).

Cortes, Corinna, Yishay Mansour, and Mehryar Mohri (2010). "Learning Bounds for Importance Weighting". In: *Advances in Neural Information Processing Systems*. Ed. by J. Lafferty, C. Williams, J. Shawe-Taylor, R. Zemel, and A. Culotta. Vol. 23. Curran Associates, Inc. (cit. on p. 40).

Cox, D. R. (1972). "Regression Models and Life-Tables". *Journal of the Royal Statistical Society. Series B (Methodological)* 34:2, pp. 187–220. JSTOR: 2985181 (cit. on pp. 11, 23, 38, 111, 215).

Cox, D. R. and D. Oakes (1984). *Analysis of Survival Data*. Chapman and Hall/CRC, Boca Raton. 212 pp. DOI: 10.1201/9781315137438 (cit. on pp. 8, 17, 50, 212, 223).

Cox, Michael A. A. and Trevor F. Cox (2008). "Multidimensional Scaling". In: *Handbook of Data Visualization*. Ed. by Chun-houh Chen, Wolfgang Härdle, and Antony Unwin. Springer Handbooks Comp.Statistics. Springer, Berlin, Heidelberg, pp. 315–347. DOI: 10.1007/978-3-540-33037-0_14 (cit. on p. 136).

Curtis, Christina et al. (18, 2012). "The Genomic and Transcriptomic Architecture of 2,000 Breast Tumours Reveals Novel Subgroups". *Nature* 486:7403, pp. 346–352. DOI: 10.1038/nature10983. pmid: 22522925 (cit. on p. 131).

Cusumano-Towner, Marco F., Feras A. Saad, Alexander K. Lew, and Vikash K. Mansinghka (8, 2019). "Gen: A General-Purpose Probabilistic Programming System with Programmable Inference". In: *Proceedings of the 40th ACM SIGPLAN Conference on Programming Language Design and Implementation*. PLDI 2019. Association for Computing Machinery, New York, NY, USA, pp. 221–236. DOI: 10.1145/3314221.3314642 (cit. on p. 187).

Cuturi, Marco, Olivier Teboul, and Jean-Philippe Vert (2019). "Differentiable Ranking and Sorting Using Optimal Transport". *Advances in Neural Information Processing Systems* 32 (cit. on pp. 62, 120).

Dabrowska, Dorota Maria (1988). "Kaplan-Meier Estimate on the Plane". *The Annals of Statistics* 16:4, pp. 1475–1489 (cit. on p. 110).

– (1989). "Uniform Consistency of the Kernel Conditional Kaplan-Meier Estimate". *Annals of Statistics* 17:3, pp. 1157–1167. DOI: 10.1214/aos/1176347261 (cit. on pp. 14, 37, 42, 44, 45, 219).

Dalalyan, Arnak S., Anatoli Juditsky, and Vladimir Spokoiny (2008). "A New Algorithm for Estimating the Effective Dimension-Reduction Subspace". *Journal of Machine Learning Research* 9, pp. 1647–1678 (cit. on pp. 140, 142, 149, 160).

Dawid, A. Philip and Monica Musio (2014). "Theory and Applications of Proper Scoring Rules". *METRON* 72:2, pp. 169–183. DOI: 10.1007/s40300-014-0039-y. arXiv: 1401.0398 (cit. on p. 62).

De Brabanter, Kris, Jos De Brabanter, Bart De Moor, and Irène Gijbels (1, 2013). "Derivative Estimation with Local Polynomial Fitting". *The Journal of Machine Learning Research* 14:1, pp. 281–301 (cit. on p. 141).

Delecroix, M. and A. C. Rosa (1, 1996). "Nonparametric Estimation of a Regression Function and Its Derivatives under an Ergodic Hypothesis". *Journal of Nonparametric Statistics* 6:4, pp. 367–382. DOI: 10.1080/10485259608832682 (cit. on p. 141).

Delyon, Bernard and François Portier (2016). "Integral Approximation by Kernel Smoothing". *Bernoulli* 22:4, pp. 2177–2208. DOI: 10.3150/15-BEJ725 (cit. on p. 48).

– (20, 2020). "Safe and Adaptive Importance Sampling: A Mixture Approach". *Annals of Statistics*. arXiv: 1903.08507 (cit. on pp. 57, 73).

Dempster, A. P., N. M. Laird, and Donald B. Rubin (1977). "Maximum Likelihood from Incomplete Data via the EM Algorithm". *Journal of the Royal Statistical Society. Series B (Methodological)* 39:1, pp. 1–38. JSTOR: 2984875 (cit. on p. 112).

Detrano, R. et al. (1, 1989). "International Application of a New Probability Algorithm for the Diagnosis of Coronary Artery Disease". *The American Journal of Cardiology* 64:5, pp. 304–310. DOI: 10.1016/0002-9149(89)90524-9. pmid: 2756873 (cit. on p. 150).

Devroye, Luc, László Györfi, and Gábor Lugosi (1996). *A Probabilistic Theory of Pattern Recognition*. Stochastic Modelling and Applied Probability. Springer-Verlag, New York. doi: 10.1007/978-1-4612-0711-5 (cit. on pp. 32, 139).

Dhariwal, Prafulla, Heewoo Jun, Christine Payne, Jong Wook Kim, Alec Radford, and Ilya Sutskever (30, 2020). *Jukebox: A Generative Model for Music*. arXiv: 2005.00341 [cs, eess, stat]. url: http://arxiv.org/abs/2005.00341 (cit. on p. 116).

Dietz, Klaus and J. A. P. Heesterbeek (1, 2002). "Daniel Bernoulli's Epidemiological Model Revisited". *Mathematical Biosciences* 180:1, pp. 1–21. doi: 10.1016/S0025-5564(02)00122-0 (cit. on p. 201).

Dinh, Laurent, David Krueger, and Yoshua Bengio (10, 2015). *NICE: Non-Linear Independent Components Estimation*. arXiv: 1410.8516 [cs]. url: http://arxiv.org/abs/1410.8516 (cit. on p. 120).

Dinh, Laurent, Jascha Sohl-Dickstein, and Samy Bengio (27, 2017). *Density Estimation Using Real NVP*. arXiv: 1605.08803 [cs, stat]. url: http://arxiv.org/abs/1605.08803 (cit. on p. 121).

Dopierre, Thomas, Christophe Gravier, and Wilfried Logerais (27, 2021). *ProtAugment: Unsupervised Diverse Short-Texts Paraphrasing for Intent Detection Meta-Learning*. arXiv: 2105.12995 [cs]. url: http://arxiv.org/abs/2105.12995 (cit. on p. 116).

Du, Yunling and Michael G. Akritas (2002). "Uniform Strong Representation of the Conditional Kaplan-Meier Process". *Mathematical Methods of Statistics* 11:2, pp. 152–182 (cit. on p. 42).

Dudley, R. M. (1992). "Frechet Differentiability, p-Variation and Uniform Donsker Classes". *Annals of Probability* 20:4, pp. 1968–1982. doi: 10.1214/aop/1176989537 (cit. on pp. 73, 92, 96).

Dupont, Emilien, Arnaud Doucet, and Yee Whye Teh (26, 2019). *Augmented Neural ODEs*. arXiv: 1904.01681 [cs, stat]. url: http://arxiv.org/abs/1904.01681 (cit. on p. 132).

Dvoretzky, A., J. Kiefer, and J. Wolfowitz (1956). "Asymptotic Minimax Character of the Sample Distribution Function and of the Classical Multinomial Estimator". *The Annals of Mathematical Statistics* 27:3, pp. 642–669. doi: 10.1214/aoms/1177728174 (cit. on p. 37).

Einmahl, Uwe and David M. Mason (1, 2000). "An Empirical Process Approach to the Uniform Consistency of Kernel-Type Function Estimators". *Journal of Theoretical Probability* 13:1, pp. 1–37. doi: 10.1023/A:1007769924157 (cit. on p. 68).

Embrechts, Paul, Claudia Klüppelberg, and Thomas Mikosch (1997). "Risk Theory". In: *Modelling Extremal Events: For Insurance and Finance*. Ed. by Paul Embrechts, Claudia Klüppelberg, and Thomas Mikosch. Applications of Mathematics. Springer, Berlin, Heidelberg, pp. 21–57. doi: 10.1007/978-3-642-33483-2_2 (cit. on pp. 3, 207).

Fan, Jianqing (1992). "Design-Adaptive Nonparametric Regression". *Journal of the American Statistical Association* 87:420, pp. 998–1004. doi: 10.1080/01621459.1992.10476255 (cit. on pp. 143, 153).

– (1993). "Local Linear Regression Smoothers and Their Minimax Efficiencies". *The Annals of Statistics* 21:1, pp. 196–216. doi: 10.1214/aos/1176349022 (cit. on p. 153).

Fan, Jianqing and Irene Gijbels (1, 1996). *Local Polynomial Modelling and Its Applications: Monographs on Statistics and Applied Probability 66*. CRC Press. 362 pp. Google Books: BM1ckQKCXP8C (cit. on p. 141).

Fernandez, T., N. Rivera, and Yee Whye Teh (2016). "Gaussian Processes for Survival Analysis". In: *NIPS* (cit. on p. 112).

Finlay, Chris, Jörn-Henrik Jacobsen, Levon Nurbekyan, and Adam M. Oberman (23, 2020). *How to Train Your Neural ODE: The World of Jacobian and Kinetic Regularization*. arXiv: 2002.02798 [cs, stat]. URL: http://arxiv.org/abs/2002.02798 (cit. on p. 132).

Fisher, R. A. (1936). "The Use of Multiple Measurements in Taxonomic Problems". *Annals of Eugenics* 7:2, pp. 179–188. DOI: 10.1111/j.1469-1809.1936.tb02137.x (cit. on p. 136).

Fleming, T. and D. Harrington (1991). "Counting Processes and Survival Analysis". In: DOI: 10.2307/2290673 (cit. on pp. 7, 34, 74, 85, 212).

Foekens, J. A. et al. (1, 2000). "The Urokinase System of Plasminogen Activation and Prognosis in 2780 Breast Cancer Patients". *Cancer Research* 60:3, pp. 636–643. pmid: 10676647 (cit. on p. 131).

Fontanilla Ramirez, Paula Victoria (30, 2020). "Role of Lamin B1 Dysregulation in Two Enabling Forces of Cancer : Genome Instability and Inflammation Le Vieillissement Une Histoire de Dommages de l'ADN, d'enveloppe Nucléaire Altérée et d'inflammation ?" These de doctorat. Université Paris-Saclay (cit. on p. 151).

Fouché, Edouard, Junpei Komiyama, and Klemens Böhm (25, 2019). "Scaling Multi-Armed Bandit Algorithms". In: *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. KDD '19. Association for Computing Machinery, New York, NY, USA, pp. 1449–1459. DOI: 10.1145/3292500.3330862 (cit. on pp. 2, 206).

Gasser, Theo and Hans-Georg Müller (1984). "Estimating Regression Functions and Their Derivatives by the Kernel Method". *Scandinavian Journal of Statistics* 11:3, pp. 171–185. JSTOR: 4615954 (cit. on p. 141).

Ge, Hong, Kai Xu, and Zoubin Ghahramani (31, 2018). "Turing: A Language for Flexible Probabilistic Inference". In: *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*. International Conference on Artificial Intelligence and Statistics. PMLR, pp. 1682–1690 (cit. on p. 187).

Gelman, Andrew (2006). "Multilevel (Hierarchical) Modeling: What It Can and Cannot Do". *Technometrics* 48:3, pp. 432–435. DOI: 10.1198/004017005000000661 (cit. on p. 186).

Gerds, Thomas A, Jan Beyersmann, Liis Starkopf, Sandra Frank, Mark J. van der Laan, and Martin Schumacher (2017). "The Kaplan-Meier Integral in the Presence of Covariates: A Review". *From Statistics to Mathematical Finance*, pp. 25–42. DOI: 10.1007/978-3-319-50986-0 (cit. on pp. 36, 37).

Gill, R (1994). "Lectures on Survival Analysis". *Lectures on Probability Theory*, pp. 1–127. DOI: 10.1007/BFb0073873 (cit. on pp. 7, 8, 109, 212, 213).

Giné, Evarist and Armelle Guillou (1, 2001). "On Consistency of Kernel Density Estimators for Randomly Censored Data: Rates Holding Uniformly over Adaptive Intervals". *Annales de l'Institut Henri Poincare (B) Probability and Statistics* 37:4, pp. 503–522. DOI: 10.1016/S0246-0203(01)01081-0 (cit. on pp. 48, 68, 162, 163).

Giné, Evarist, Vladimir Koltchinskii, and Joel Zinn (2004). "Weighted Uniform Consistency of Kernel Density Estimators". *Annals of Probability* 32, 3B, pp. 2570–2605. DOI: 10.1214/009117904000000063 (cit. on p. 43).

Giné, Evarist and Hailin Sang (1, 2010). "Uniform Asymptotics for Kernel Density Estimators with Variable Bandwidths". *Journal of Nonparametric Statistics* 22:6, pp. 773–795. DOI: 10.1080/10485250903483331 (cit. on p. 68).

Goldstein, Mark, Xintian Han, Aahlad Puli, Thomas Wies, Adler Perotte, and Rajesh Ranganath (2021). "Inverse-Weighted Survival Games". In: NeurIPS (cit. on p. 64).

Gorban, Alexander N., Balázs Kégl, Donald C. Wunsch, and Andrei Zinovyev, eds. (2008). *Principal Manifolds for Data Visualization and Dimension Reduction*. Lecture Notes in Computational Science and Engineering. Springer-Verlag, Berlin Heidelberg. DOI: 10.1007/978-3-540-73750-6 (cit. on p. 136).

Groha, Stefan, Sebastian M. Schmon, and Alexander Gusev (8, 2020). *A General Framework for Survival Analysis and Multi-State Modelling*. arXiv: 2006.04893 [cs, stat]. URL: http://arxiv.org/abs/2006.04893 (cit. on pp. 113, 114).

Grossman, Robert L. et al. (22, 2016). "Toward a Shared Vision for Cancer Genomic Data". *New England Journal of Medicine* 375:12, pp. 1109–1112. DOI: 10.1056/NEJMp1607591. pmid: 27653561 (cit. on p. 60).

Györfi, László, Michael Kohler, Adam Krzyzak, and Harro Walk (2002). *A Distribution-Free Theory of Nonparametric Regression*. Springer Series in Statistics. Springer-Verlag, New York. DOI: 10.1007/b97848 (cit. on p. 32).

Hagan, Patrick, Andrew Lesniewski, and Diana Woodward (2015). "Probability Distribution in the SABR Model of Stochastic Volatility". In: *Large Deviations and Asymptotic Methods in Finance*. Ed. by Peter K. Friz, Jim Gatheral, Archil Gulisashvili, Antoine Jacquier, and Josef Teichmann. Springer Proceedings in Mathematics & Statistics. Springer International Publishing, Cham, pp. 1–35. DOI: 10.1007/978-3-319-11605-1_1 (cit. on p. 117).

Halley, Edmond (1, 1693). "An Estimate of the Degrees of the Mortality of Mankind; Drawn from Curious Tables of the Births and Funerals at the City of Breslaw; with an Attempt to Ascertain the Price of Annuities upon Lives". *Philosophical Transactions of the Royal Society of London* 17:196, pp. 596–610. DOI: 10.1098/rstl.1693.0007 (cit. on p. 202).

Harrell, Frank E., Robert M. Califf, David B. Pryor, Kerry L. Lee, and Robert A. Rosati (14, 1982). "Evaluating the Yield of Medical Tests". *JAMA* 247:18, pp. 2543–2546. DOI: 10.1001/jama.1982.03320430047030 (cit. on p. 129).

Harrell, Frank E., Kerry L. Lee, and Daniel B. Mark (1996). "Multivariable Prognostic Models: Issues in Developing Models, Evaluating Assumptions and Adequacy, and Measuring and Reducing Errors". *Statistics in Medicine* 15:4, pp. 361–387. DOI: 10.1002/(SICI)1097-0258(19960229)15:4<361::AID-SIM168>3.0.CO;2-4 (cit. on p. 129).

Harvey, Félix G., Mike Yurick, Derek Nowrouzezahrai, and Christopher Pal (8, 2020). "Robust Motion In-Betweening". *ACM Transactions on Graphics* 39:4. DOI: 10.1145/3386569.3392480 (cit. on p. 116).

Hastie, Trevor and Werner Stuetzle (1, 1989). "Principal Curves". *Journal of the American Statistical Association* 84:406, pp. 502–516. DOI: 10.1080/01621459.1989.10478797 (cit. on p. 135).

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman (2008). *The Elements of Statistical Learning*. 3rd ed. (cit. on p. 32).

Hastie, Trevor, Robert Tibshirani, and Martin J. Wainwright (2015). *Statistical Learning with Sparsity: The Lasso and Generalizations*. Monographs on Statistics and Applied Probability. Chapman and Hall/CRC (cit. on pp. 147, 172).

Henry-Labordere, Pierre (21, 2019). *Generative Models for Financial Data*. SSRN Scholarly Paper ID 3408007. Rochester, NY: Social Science Research Network. DOI: 10.2139/ssrn.3408007 (cit. on p. 117).

Heston, Steven L. (1, 1993). "A Closed-Form Solution for Options with Stochastic Volatility with Applications to Bond and Currency Options". *The Review of Financial Studies* 6:2, pp. 327–343. DOI: 10.1093/rfs/6.2.327 (cit. on p. 117).

Higgins, Irina et al. (2017). "Beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework". *ICLR* (cit. on p. 156).

Hinton, Geoffrey E and Sam T. Roweis (2003). "Stochastic Neighbor Embedding". In: *Advances in Neural Information Processing Systems*. Vol. 15. MIT Press (cit. on p. 136).

Hirsch, Morris W., Stephen Smale, and Robert L. Devaney (2013). *Differential Equations, Dynamical Systems, and an Introduction to Chaos*. Elsevier. DOI: 10.1016/C2009-0-61160-0 (cit. on p. 122).

Holden, Daniel, Oussama Kanoun, Maksym Perepichka, and Tiberiu Popa (8, 2020). "Learned Motion Matching". *ACM Transactions on Graphics* 39:4. DOI: 10.1145/3386569.3392440 (cit. on p. 116).

Hosmer, David, Stanley Lemeshow, and Susanne May (1, 2000). "Applied Survival Analysis: Regression Modeling of Time to Event Data". *Journal of the American Statistical Association* 95. DOI: 10.2307/2669422 (cit. on p. 130).

Hothorn, Torsten, Peter Bühlmann, Sandrine Dudoit, Annette Molinaro, and Mark J. van der Laan (2006). "Survival Ensembles". *Biostatistics* 7:3, pp. 355–373. DOI: 10.1093/biostatistics/kxj011 (cit. on p. 60).

Hristache, Marian, Anatoli Juditsky, Jörg Polzehl, and Vladimir Spokoiny (2001). "Structure Adaptive Approach for Dimension Reduction". *Annals of Statistics* 29:6, pp. 1537–1566. DOI: 10.1214/aos/1015345954 (cit. on p. 139).

Hristache, Marian, Anatoli Juditsky, and Vladimir Spokoiny (1998). *Direct Estimation of the Index Coefficients in a Single-Index Model*. report. INRIA (cit. on p. 139).

Hyvonen, Ville et al. (2016). "Fast Nearest Neighbor Search through Sparse Random Projections and Voting". In: *2016 IEEE International Conference on Big Data (Big Data)*. 2016 IEEE International Conference on Big Data (Big Data). IEEE, Washington DC,USA, pp. 881–888. DOI: 10.1109/BigData.2016.7840682 (cit. on p. 152).

Innes, Mike et al. (18, 2019). *A Differentiable Programming System to Bridge Machine Learning and Scientific Computing*. arXiv: 1907.07587 [cs]. URL: http://arxiv.org/abs/1907.07587 (cit. on pp. 119, 125, 128, 152).

Ishwaran, Hemant and Udaya B. Kogalur (2007). "Random Survival Forests for R". *R News* 7:2, pp. 25–31 (cit. on pp. 17, 60, 149, 223).

Ishwaran, Hemant, Udaya B. Kogalur, Eugene H. Blackstone, and Michael S. Lauer (2008). "Random Survival Forests". *The Annals of Applied Statistics* 2:3, pp. 841–860. DOI: 10.1214/08-AOAS169. pmid: 261057900003 (cit. on pp. 52, 60, 130, 132).

Jiang, Heinrich (17, 2019). "Non-Asymptotic Uniform Rates of Consistency for k-NN Regression". *Proceedings of the AAAI Conference on Artificial Intelligence* AAAI Technical Track: Machine Learning, Vol 33 No 01: AAAI-19, IAAI-19, EAAI-20. arXiv: 1707.06261 (cit. on pp. 141, 145–147, 159).

Johnson, William B., Joram Lindenstrauss, and Gideon Schechtman (1, 1986). "Extensions of Lipschitz Maps into Banach Spaces". *Israel Journal of Mathematics* 54:2, pp. 129–138. doi: 10.1007/BF02764938 (cit. on p. 136).

Kalivas, John H. (1997). "Two Data Sets of near Infrared Spectra". *Chemometrics and Intelligent Laboratory Systems* 37:2, pp. 255–259. doi: 10.1016/S0169-7439(97)00038-5 (cit. on p. 150).

Kaplan, E. L. and Paul Meier (1958). "Nonparametric Estimation from Incomplete Observations". *Journal of the American Statistical Association* 53:282, pp. 457–481. doi: 10.2307/2281868. JSTOR: 2281868 (cit. on pp. 9, 38, 107, 213).

Karras, Tero, Samuli Laine, and Timo Aila (29, 2019). *A Style-Based Generator Architecture for Generative Adversarial Networks*. arXiv: 1812.04948 [cs, stat]. url: http://arxiv.org/abs/1812.04948 (cit. on p. 116).

Karras, Tero, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila (23, 2020). *Analyzing and Improving the Image Quality of StyleGAN*. arXiv: 1912.04958 [cs, eess, stat]. url: http://arxiv.org/abs/1912.04958 (cit. on p. 116).

Katzman, Jared, Uri Shaham, Jonathan Bates, Alexander Cloninger, Tingting Jiang, and Yuval Kluger (2018). "DeepSurv: Personalized Treatment Recommender System Using A Cox Proportional Hazards Deep Neural Network". *BMC Medical Research Methodology* 18:1, p. 24. doi: 10.1186/s12874-018-0482-1. arXiv: 1606.00931 (cit. on pp. 111, 130, 132).

Khalil, Hassan K. (2002). *Nonlinear Systems*. Prentice Hall. 750 pp. Google Books: t_d1QgAACAAJ (cit. on p. 122).

Kim, Yongdai and Jinseog Kim (4, 2004). "Gradient LASSO for Feature Selection". In: *Proceedings of the Twenty-First International Conference on Machine Learning*. ICML '04. Association for Computing Machinery, New York, NY, USA, p. 60. doi: 10.1145/1015330.1015364 (cit. on p. 137).

Kingma, Diederik P. and Prafulla Dhariwal (10, 2018). *Glow: Generative Flow with Invertible 1x1 Convolutions*. arXiv: 1807.03039 [cs, stat]. url: http://arxiv.org/abs/1807.03039 (cit. on p. 121).

Kingma, Diederik P. and Max Welling (2014). "Auto-Encoding Variational Bayes". *ICLR*. arXiv: 1312.6114 (cit. on pp. 118, 136, 157).

– (2019). "An Introduction to Variational Autoencoders". *Foundations and Trends® in Machine Learning* 12:4, pp. 307–392. doi: 10.1561/2200000056. arXiv: 1906.02691 (cit. on pp. 118, 136).

Klambauer, Günter, Thomas Unterthiner, Andreas Mayr, and Sepp Hochreiter (7, 2017). *Self-Normalizing Neural Networks*. Version 5. arXiv: 1706.02515 [cs, stat]. url: http://arxiv.org/abs/1706.02515 (cit. on p. 131).

Klein, John P. and Melvin L. Moeschberger (2003). *Survival Analysis: Techniques for Censored and Truncated Data*. 2nd ed. Statistics for Biology and Health. Springer-Verlag, New York. doi: 10.1007/b97377 (cit. on pp. 8, 24, 212).

Knaus, William A. et al. (1, 1995). "The SUPPORT Prognostic Model: Objective Estimates of Survival for Seriously Ill Hospitalized Adults". *Annals of Internal Medicine* 122:3, pp. 191–203. DOI: 10.7326/0003-4819-122-3-199502010-00007 (cit. on p. 130).

Kohler, Michael, Kinga Máthé, and Márta Pintér (2002). "Prediction from Randomly Right Censored Data". *Journal of Multivariate Analysis* 80:1, pp. 73–100. DOI: 10.1006/jmva.2000.1973 (cit. on p. 39).

Kolmogorov, Andrey Nikolaevich (1955). "Estimates of the Minimal Number of Elements of e Nets in Various Functional Classes and Their Applications to the Question of Representing Functions of Several Variables as Superpositions of Functions of a Lesser Number of Variables". *Uspekhi Mat. Nauk* 10, pp. 192–193 (cit. on p. 31).

– (1956). "Certain Asymptotic Characteristics of Completely Bounded Metric Spaces." *Doklady Akademii Nauk SSSR* 108:3, pp. 385–388 (cit. on p. 31).

Kpotufe, Samory (2011). "K-NN Regression Adapts to Local Intrinsic Dimension". In: *Advances in Neural Information Processing Systems 24*. Ed. by J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, and K. Q. Weinberger. Curran Associates, Inc., pp. 729–737 (cit. on pp. 141, 145, 147).

Kvamme, Håvard, Ornulf Borgan, and Ida Scheel (2019). "Time-to-Event Prediction with Neural Networks and Cox Regression". *Journal of Machine Learning Research* 20:129, pp. 1–30 (cit. on pp. 63, 111).

Lecué, Guillaume and Shahar Mendelson (17, 2016). *Learning Subgaussian Classes : Upper and Minimax Bounds*. arXiv: 1305.4825 [math, stat]. URL: http://arxiv.org/abs/1305.4825 (cit. on p. 32).

Lee, C., J. Yoon, and M. v d Schaar (2020). "Dynamic-DeepHit: A Deep Learning Approach for Dynamic Survival Analysis With Competing Risks Based on Longitudinal Data". *IEEE Transactions on Biomedical Engineering* 67:1, pp. 122–133. DOI: 10.1109/TBME.2019.2909027 (cit. on pp. 17, 114, 222).

Lee, C., W. Zame, Jinsung Yoon, and M. V. D. Schaar (2018). "DeepHit: A Deep Learning Approach to Survival Analysis With Competing Risks". In: *AAAI* (cit. on pp. 17, 114, 222).

Lewis, Mike et al. (2020). "BART: Denoising Sequence-to-Sequence Pre-Training for Natural Language Generation, Translation, and Comprehension". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. ACL 2020. Association for Computational Linguistics, Online, pp. 7871–7880. DOI: 10.18653/v1/2020.acl-main.703 (cit. on p. 116).

Lopez, Olivier (2011). "Nonparametric Estimation of the Multivariate Distribution Function in a Censored Regression Model with Applications". *Communications in Statistics - Theory and Methods* 40:15, pp. 2639–2660. DOI: 10.1080/03610926.2010.489175 (cit. on pp. 38, 39).

Lopez, Olivier, Valentin Patilea, and Ingrid van Keilegom (2013). "Single Index Regression Models in the Presence of Censoring Depending on the Covariates". *Bernoulli* 19:3, pp. 721–747. DOI: 10.3150/12-BEJ464 (cit. on pp. 38, 39).

Lu, Haiping, Konstantinos N. Plataniotis, and Anastasios N. Venetsanopoulos (1, 2011). "A Survey of Multilinear Subspace Learning for Tensor Data". *Pattern Recognition* 44:7, pp. 1540–1551. DOI: 10.1016/j.patcog.2011.01.004 (cit. on p. 135).

Lugosi, Gábor and Shahar Mendelson (2016). "Risk Minimization by Median-of-Means Tournaments". *Journal of the European Mathematical Society* 22:3. arXiv: 1608.00757 (cit. on p. 32).

Major, Péter (2006). "An Estimate on the Supremum of a Nice Class of Stochastic Integrals and U-Statistics". *Probability Theory and Related Fields* 134:3, pp. 489–537. DOI: 10.1007/s00440-005-0440-9 (cit. on p. 68).

Malkov, Y. A. and D. A. Yashunin (2020). "Efficient and Robust Approximate Nearest Neighbor Search Using Hierarchical Navigable Small World Graphs". *IEEE Transactions on Pattern Analysis and Machine Intelligence* 42:4, pp. 824–836. DOI: 10.1109/TPAMI.2018.2889473 (cit. on p. 148).

Mann, Nancy R., Ray E. Schafer, and Nozer D. Singpurwalla (1974). *Methods for Statistical Analysis of Reliability and Life Data*. Wiley. 586 pp. Google Books: yfdTAAAAMAAJ (cit. on p. 33).

Margossian, Charles C. (2019). "A Review of Automatic Differentiation and Its Efficient Implementation". *WIREs Data Mining and Knowledge Discovery* 9:4. DOI: 10.1002/WIDM.1305. arXiv: 1811.05031 (cit. on p. 119).

Markowitz, Harry (1952a). "Portfolio Selection". *The Journal of Finance* 7:1, pp. 77–91. DOI: 10.2307/2975974. JSTOR: 2975974 (cit. on p. 179).

– (1, 1952b). "The Utility of Wealth". *Journal of Political Economy* 60:2, pp. 151–158. DOI: 10.1086/257177 (cit. on p. 179).

Marti, Gautier (2020). "CorrGAN: Sampling Realistic Financial Correlation Matrices Using Generative Adversarial Networks". *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 8459–8463. DOI: 10.1109/ICASSP40776.2020.9053276. arXiv: 1910.09504 (cit. on p. 117).

Marti, Gautier, Victor Goubet, and Frank Nielsen (2021). *cCorrGAN: Conditional Correlation GAN for Learning Empirical Conditional Distributions in the Elliptope*. DOI: 10.1007/978-3-030-80209-7_66. arXiv: 2107.10606 [cs, q-fin]. URL: http://arxiv.org/abs/2107.10606 (cit. on p. 117).

Martinsson, Egil (2016). "WTTE-RNN : Weibull Time To Event Recurrent Neural Network" (cit. on pp. 114, 115).

Massart, Pascal (2007). *Concentration Inequalities and Model Selection: Ecole d'Eté de Probabilités de Saint-Flour XXXIII - 2003*. Ed. by Jean Picard. École d'Été de Probabilités de Saint-Flour. Springer-Verlag, Berlin Heidelberg. DOI: 10.1007/978-3-540-48503-2 (cit. on p. 32).

McInnes, Leland, John Healy, and James Melville (17, 2020). *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction*. arXiv: 1802.03426 [cs, stat]. URL: http://arxiv.org/abs/1802.03426 (cit. on p. 136).

Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean (2013). "Efficient Estimation of Word Representations in Vector Space". *Arxiv*, pp. 1–12. DOI: 10.1162/153244303322533223 (cit. on p. 136).

Miller, Rupert and Jerry Halpern (1982). "Regression with Censored Data". *Biometrika* 69:3, pp. 521–531. DOI: 10.2307/2335987. JSTOR: 2335987 (cit. on p. 123).

Miscouridou, Xenia, Adler Perotte, Noemie Elhadad, and Rajesh Ranganath (29, 2018). "Deep Survival Analysis: Nonparametrics and Missingness". In: *Machine Learning for*

Healthcare Conference. Machine Learning for Healthcare Conference. PMLR, pp. 244–256 (cit. on p. 115).

Molinaro, Annette M., Sandrine Dudoit, and Mark J. van der Laan (2004). "Tree-Based Multivariate Regression and Density Estimation with Right-Censored Data". *Journal of Multivariate Analysis* 90, 1 SPEC. ISS. Pp. 154–177. DOI: 10.1016/j.jmva.2004.02.003 (cit. on p. 38).

Mukherjee, Sayan and Qiang Wu (1, 2006). "Estimation of Gradients and Coordinate Covariation in Classification". *The Journal of Machine Learning Research* 7, pp. 2481–2514 (cit. on p. 139).

Mukherjee, Sayan and Ding-Xuan Zhou (1, 2006). "Learning Coordinate Covariances via Gradients". *The Journal of Machine Learning Research* 7, pp. 519–549 (cit. on pp. 139, 141).

Nagpal, Chirag, Steve Yadlowsky, Negar Rostamzadeh, and Katherine Heller (16, 2021). *Deep Cox Mixtures for Survival Regression*. arXiv: 2101.06536 [cs, stat]. URL: http://arxiv.org/abs/2101.06536 (cit. on pp. 17, 111, 112, 222).

Nelson, Wayne (1, 1969). "Hazard Plotting for Incomplete Failure Data". *Journal of Quality Technology* 1:1, pp. 27–52. DOI: 10.1080/00224065.1969.11980344 (cit. on pp. 9, 110, 214).

– (1, 1972). "Theory and Applications of Hazard Plotting for Censored Failure Data". *Technometrics* 14:4, pp. 945–966. DOI: 10.1080/00401706.1972.10488991 (cit. on p. 110).

Nerney, J. S. Mac (1963). "Integral Equations and Semigroups". *Illinois Journal of Mathematics* 7:1, pp. 148–173. DOI: 10.1215/ijm/1255637489 (cit. on p. 109).

Nesterov, Yurii and Vladimir Spokoiny (1, 2017). "Random Gradient-Free Minimization of Convex Functions". *Foundations of Computational Mathematics* 17:2, pp. 527–566. DOI: 10.1007/s10208-015-9296-2 (cit. on p. 140).

Nichol, Alexander Quinn and Prafulla Dhariwal (1, 2021). "Improved Denoising Diffusion Probabilistic Models". In: *Proceedings of the 38th International Conference on Machine Learning*. International Conference on Machine Learning. PMLR, pp. 8162–8171 (cit. on p. 187).

Nolan, Deborah and David Pollard (1987). "U-Processes: Rates of Convergence". *Annals of Statistics* 15:2, pp. 780–799. DOI: 10.1214/aos/1176350374 (cit. on pp. 43, 48, 68, 70, 93).

Oliva, Junier B. et al. (23, 2018). *Transformation Autoregressive Networks*. arXiv: 1801.09819 [stat]. URL: http://arxiv.org/abs/1801.09819 (cit. on p. 121).

Oord, Aaron van den et al. (19, 2016). *WaveNet: A Generative Model for Raw Audio*. arXiv: 1609.03499 [cs]. URL: http://arxiv.org/abs/1609.03499 (cit. on p. 116).

Pan, Sinno Jialin and Qiang Yang (2010). "A Survey on Transfer Learning". *IEEE Transactions on Knowledge and Data Engineering* 22:10, pp. 1345–1359. DOI: 10.1109/TKDE.2009.191 (cit. on p. 40).

Papa, Guillaume, Aurélien Bellet, and Stéphan Clémençon (2016). "On Graph Reconstruction via Empirical Risk Minimization: Fast Learning Rates and Scalability". In: *Advances in Neural Information Processing Systems 29*. Ed. by D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett. Curran Associates, Inc., pp. 694–702 (cit. on p. 48).

Papamakarios, George, Theo Pavlakou, and Iain Murray (14, 2018). *Masked Autoregressive Flow for Density Estimation*. arXiv: 1705.07057 [cs, stat]. URL: http://arxiv.org/abs/1705.07057 (cit. on p. 121).

Pearson, Karl (1, 1901). "On Lines and Planes of Closest Fit to Systems of Points in Space". *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 2:11, pp. 559–572. doi: 10.1080/14786440109462720 (cit. on p. 135).

Pedregosa, Fabian et al. (2011). "Scikit-Learn: Machine Learning in Python". *Journal of Machine Learning Research* 12, pp. 2825–2830 (cit. on p. 60).

Peña, Victor de la and Evarist Giné (1999). *Decoupling: From Dependence to Independence*. Probability and Its Applications. Springer-Verlag, New York. doi: 10.1007/978-1-4612-0537-1 (cit. on p. 48).

Pennington, Jeffrey, Richard Socher, and Christopher Manning (2014). "Glove: Global Vectors for Word Representation". In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Association for Computational Linguistics, Doha, Qatar, pp. 1532–1543. doi: 10.3115/v1/D14-1162 (cit. on p. 136).

Pollard, Andrew J. and Else M. Bijker (2021). "A Guide to Vaccinology: From Basic Principles to New Developments". *Nature Reviews Immunology* 21:2, 2, pp. 83–100. doi: 10.1038/s41577-020-00479-7 (cit. on pp. 4, 208).

Pölsterl, Sebastian (2020). "Scikit-Survival: A Library for Time-to-Event Analysis Built on Top of Scikit-Learn". *Journal of Machine Learning Research* 21:212, pp. 1–6 (cit. on p. 60).

Pölsterl, Sebastian, Nassir Navab, and Amin Katouzian (2015). "Fast Training of Support Vector Machines for Survival Analysis". In: *Machine Learning and Knowledge Discovery in Databases*. Ed. by Annalisa Appice, Pedro Pereira Rodrigues, Vítor Santos Costa, João Gama, Alípio Jorge, and Carlos Soares. Lecture Notes in Computer Science. Springer International Publishing, pp. 243–259 (cit. on p. 60).

– (2016). "An Efficient Training Algorithm for Kernel Survival Support Vector Machines". In: *CML PKDD MLLS 2016*. CML PKDD MLLS. arXiv: 1611.07054 (cit. on p. 60).

Portier, François and Bernard Delyon (2, 2014). "Bootstrap Testing of the Rank of a Matrix via Least-Squared Constrained Estimation". *Journal of the American Statistical Association* 109:505, pp. 160–172. doi: 10.1080/01621459.2013.847841 (cit. on p. 150).

Portier, François and Johan Segers (2018). "On the Weak Convergence of the Empirical Conditional Copula under a Simplifying Assumption". *Journal of Multivariate Analysis* 166:C, pp. 160–181 (cit. on p. 70).

Rackauckas, Christopher, Yingbo Ma, Vaibhav Dixit, et al. (5, 2018). *A Comparison of Automatic Differentiation and Continuous Sensitivity Analysis for Derivatives of Differential Equation Solutions*. arXiv: 1812.01892 [cs]. url: http://arxiv.org/abs/1812.01892 (cit. on pp. 17, 125, 222).

Rackauckas, Christopher, Yingbo Ma, Julius Martensen, et al. (6, 2020). *Universal Differential Equations for Scientific Machine Learning*. arXiv: 2001.04385 [cs, math, q-bio, stat]. url: http://arxiv.org/abs/2001.04385 (cit. on p. 120).

Rackauckas, Christopher and Qing Nie (2017a). "Adaptive Methods for Stochastic Differential Equations via Natural Embeddings and Rejection Sampling with Memory". *Discrete and Continuous Dynamical Systems. Series B* 22:7, pp. 2731–2761. doi: 10.3934/dcdsb.2017133. pmid: 29527134 (cit. on p. 120).

Rackauckas, Christopher and Qing Nie (25, 2017b). "DifferentialEquations.Jl – A Performant and Feature-Rich Ecosystem for Solving Differential Equations in Julia". *Journal of Open Research Software* 5:1, 1, p. 15. DOI: 10.5334/jors.151 (cit. on p. 128).

Rahimi, Ali and Benjamin Recht (2008). "Random Features for Large-Scale Kernel Machines". In: *Advances in Neural Information Processing Systems*. Vol. 20. Curran Associates, Inc. (cit. on p. 136).

Ran, Yongyi, Xin Zhou, Pengfeng Lin, Yonggang Wen, and Ruilong Deng (12, 2019). *A Survey of Predictive Maintenance: Systems, Purposes and Approaches*. arXiv: 1912.07383 [cs, eess]. URL: http://arxiv.org/abs/1912.07383 (cit. on pp. 2, 206).

Ranganath, Rajesh, Adler Perotte, Noémie Elhadad, and David Blei (10, 2016). "Deep Survival Analysis". In: *Machine Learning for Healthcare Conference*. Machine Learning for Healthcare Conference. PMLR, pp. 101–114 (cit. on pp. 114, 115).

Raudenbush, Stephen W. (1, 1988). "Educational Applications of Hierarchical Linear Models: A Review". *Journal of Educational Statistics* 13:2, pp. 85–116. DOI: 10.3102/10769986013002085 (cit. on p. 186).

Rebuffel, Clément, Laure Soulier, Geoffrey Scoutheeten, and Patrick Gallinari (2020). "A Hierarchical Model for Data-to-Text Generation". In: *Advances in Information Retrieval*. Ed. by Joemon M. Jose et al. Lecture Notes in Computer Science. Springer International Publishing, Cham, pp. 65–80. DOI: 10.1007/978-3-030-45439-5_5 (cit. on p. 116).

Rezende, Danilo and Shakir Mohamed (1, 2015). "Variational Inference with Normalizing Flows". In: *International Conference on Machine Learning*. International Conference on Machine Learning. PMLR, pp. 1530–1538 (cit. on pp. 17, 107, 120, 222).

Robins, James M. and Dianne M. Finkelstein (2000). "Correcting for Noncompliance and Dependent Censoring in an AIDS Clinical Trial with Inverse Probability of Censoring (IPCW) Log-Rank Tests". *Biometrics* 56, pp. 779–788 (cit. on p. 149).

Rockafellar, R. Tyrrell and Stanislav Uryasev (2002). "Conditional Value-at-Risk for General Loss Distributions". *Journal of Banking & Finance* 26:7, pp. 1443–1471 (cit. on p. 180).

Rotnitzky, Andrea and James M. Robins (1992). "Recovery of Information and Adjustment for Dependent Censoring Using Surrogate Markers". *AIDS Epidemiology* 88:424, p. 1473. DOI: 10.1007/978-1-4757-1229-2_14. JSTOR: 2291304 (cit. on p. 38).

Roweis, Sam T. and Lawrence K. Saul (22, 2000). "Nonlinear Dimensionality Reduction by Locally Linear Embedding". *Science* (cit. on p. 136).

Royston, Patrick and Mahesh K. B. Parmar (2011). "The Use of Restricted Mean Survival Time to Estimate the Treatment Effect in Randomized Clinical Trials When the Proportional Hazards Assumption Is in Doubt". *Statistics in Medicine* 30:19, pp. 2409–2421. DOI: 10.1002/sim.4274 (cit. on p. 53).

Rubin, Daniel and Mark J. van der Laan (12, 2007). "A Doubly Robust Censoring Unbiased Transformation". *The International Journal of Biostatistics* 3:1. DOI: 10.2202/1557-4679.1052 (cit. on p. 38).

Rubin, Donald B. (1976). "Inference and Missing Data". *Biometrika* 63:3, pp. 581–592. DOI: 10.2307/2335739. JSTOR: 2335739 (cit. on p. 35).

Ruiz-Hernández, Diego, Jesús M. Pinar-Pérez, and David Delgado-Gómez (1, 2020). "Multi-Machine Preventive Maintenance Scheduling with Imperfect Interventions: A Restless

Bandit Approach". *Computers & Operations Research* 119, p. 104927. DOI: 10.1016/j.cor.2020.104927 (cit. on pp. 2, 206).

Salakhutdinov, Ruslan and Geoffrey E Hinton (11, 2007). "Learning a Nonlinear Embedding by Preserving Class Neighbourhood Structure". In: *Artificial Intelligence and Statistics*. Artificial Intelligence and Statistics. PMLR, pp. 412–419 (cit. on p. 137).

Salimans, Tim, Andrej Karpathy, Xi Chen, and Diederik P. Kingma (19, 2017). *PixelCNN++: Improving the PixelCNN with Discretized Logistic Mixture Likelihood and Other Modifications*. arXiv: 1701.05517 [cs, stat]. URL: http://arxiv.org/abs/1701.05517 (cit. on p. 158).

Salvatier, John, Thomas V. Wiecki, and Christopher Fonnesbeck (6, 2016). "Probabilistic Programming in Python Using PyMC3". *PeerJ Computer Science* 2, e55. DOI: 10.7717/peerj-cs.55 (cit. on p. 187).

Sandfort, Veit, Ke Yan, Perry J. Pickhardt, and Ronald M. Summers (15, 2019). "Data Augmentation Using Generative Adversarial Networks (CycleGAN) to Improve Generalizability in CT Segmentation Tasks". *Scientific Reports* 9:1, 1, p. 16884. DOI: 10.1038/s41598-019-52737-x (cit. on p. 116).

Scaman, Kevin and Aladin Virmaux (2018). "Lipschitz Regularity of Deep Neural Networks: Analysis and Efficient Estimation". In: *NeurIPS* (cit. on p. 122).

Schölkopf, Bernhard, Alexander Smola, and Klaus-Robert Müller (1, 1998). "Nonlinear Component Analysis as a Kernel Eigenvalue Problem". *Neural Computation* 10:5, pp. 1299–1319. DOI: 10.1162/089976698300017467 (cit. on p. 135).

Schumacher, M. et al. (1994). "Randomized 2 x 2 Trial Evaluating Hormonal Treatment and the Duration of Chemotherapy in Node-Positive Breast Cancer Patients. German Breast Cancer Study Group". *Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology* 12:10, pp. 2086–2093. DOI: 10.1200/JCO.1994.12.10.2086. pmid: 7931478 (cit. on p. 131).

Sheth, Rishit and Nicolo Fusi (27, 2019). *Feature Gradients: Scalable Feature Selection via Discrete Relaxation*. arXiv: 1908.10382 [cs, stat]. URL: http://arxiv.org/abs/1908.10382 (cit. on p. 137).

Shimokawa, Asanao, Yohei Kawasaki, and Etsuo Miyaoka (2015). "Comparison of Splitting Methods on Survival Tree". *International Journal of Biostatistics* 11:1, pp. 175–188. DOI: 10.1515/ijb-2014-0029. pmid: 25849798 (cit. on p. 149).

Shorack, Galen R. and Jon A. Wellner (1, 2009). *Empirical Processes with Applications to Statistics*. Classics in Applied Mathematics. Society for Industrial and Applied Mathematics. 991 pp. DOI: 10.1137/1.9780898719017 (cit. on p. 74).

Shreve, Steven (2004). *Stochastic Calculus for Finance II: Continuous-Time Models*. Springer Finance Textbooks, Stochastic Calculus for Finance. Springer-Verlag, New York (cit. on pp. 3, 207).

Steingrimsson, Jon Arni, Liqun Diao, Annette M. Molinaro, and Robert L. Strawderman (10, 2016). "Doubly Robust Survival Trees: Doubly Robust Survival Trees". *Statistics in Medicine* 35:20, pp. 3595–3612. DOI: 10.1002/sim.6949 (cit. on p. 38).

Steingrimsson, Jon Arni, Liqun Diao, and Robert L. Strawderman (2, 2019). "Censoring Unbiased Regression Trees and Ensembles". *Journal of the American Statistical Association* 114:525, pp. 370–383. DOI: 10.1080/01621459.2017.1407775 (cit. on p. 38).

Steingrimsson, Jon Arni and Samantha Morrison (2020). "Deep Learning for Survival Outcomes". *Statistics in Medicine* 39:17, pp. 2339–2349. DOI: 10.1002/sim.8542 (cit. on p. 53).

Stephane, Fotso (2019). *PySurvival: Open Source Package for Survival Analysis Modeling* (cit. on p. 130).

Stone, Charles J. (1982). "Optimal Global Rates of Convergence for Nonparametric Regression". *Annals of Statistics* 10:4, pp. 1040–1053. DOI: 10.1214/aos/1176345969 (cit. on p. 147).

Strassen, Volker (1, 1969). "Gaussian Elimination Is Not Optimal". *Numerische Mathematik* 13:4, pp. 354–356. DOI: 10.1007/BF02165411 (cit. on p. 120).

Street, W. N., W. H. Wolberg, and O. L. Mangasarian (29, 1993). "Nuclear Feature Extraction for Breast Tumor Diagnosis". In: IS&T/SPIE's Symposium on Electronic Imaging: Science and Technology. Ed. by Raj S. Acharya and Dmitry B. Goldgof. San Jose, CA, pp. 861–870. DOI: 10.1117/12.148698 (cit. on p. 150).

Stute, Winfried (1993a). "Almost Sure Representations of the Product-Limit Estimator for Truncated Data". *The Annals of Statistics* 21:1, pp. 146–156. DOI: 10.1214/08-AOS620 (cit. on pp. 14, 37, 219).

– (1993b). "Consistent Estimation under Random Censorship When Covariables Are Present". *Journal of Multivariate Analysis* 45:1. DOI: 10.1006/jmva.1993.1028 (cit. on pp. 14, 37, 39, 219).

– (1995a). "The Central Limit Theorem Under Random Censorship". *Annals of Statistics* 23:2, pp. 422–439. DOI: 10.1214/aos/1176324528 (cit. on pp. 14, 37, 51, 219).

– (1995b). "The Statistical Analysis of Kaplan-Meier Integrals". *Analysis of censored data* 27, pp. 231–254. DOI: 10.1214/lnms/1215452223 (cit. on pp. 14, 37, 219).

– (1996). "Distributional Convergence under Random Censorship When Covariables Are Present". *Scandinavian Journal of Statistics* 23:4, pp. 461–471 (cit. on pp. 14, 37, 39, 219).

– (2003). "Kaplan-Meier Integrals". *Handbook of Statistics* 23:03, pp. 87–104. DOI: 10.1016/S0169-7161(03)23005-4 (cit. on pp. 14, 219).

Stute, Winfried and J.-L. Wang (1993). "The Strong Law under Random Censorship". *The Annals of Statistics* 21:3, pp. 1591–1607. DOI: 10.1214/aos/1176349273 (cit. on pp. 14, 37, 219).

Sutskever, Ilya, Oriol Vinyals, and Quoc V Le (2014). "Sequence to Sequence Learning with Neural Networks". In: *Advances in Neural Information Processing Systems*. Vol. 27. Curran Associates, Inc. (cit. on p. 116).

Takahashi, Shuntaro, Yu Chen, and Kumiko Tanaka-Ishii (2019). "Modeling Financial Time-Series with Generative Adversarial Networks". *Physica A: Statistical Mechanics and its Applications* 527, p. 121261. DOI: 10.1016/j.physa.2019.121261 (cit. on p. 117).

Talagrand, Michel (1, 1996). "New Concentration Inequalities in Product Spaces". *Inventiones mathematicae* 126:3, pp. 505–563. DOI: 10.1007/s002220050108 (cit. on p. 163).

Tikhomirov, Vladimir Mikhailovich (1957). "On the ⊠-Entropy of Certain Classes of Analytic Functions." *Dokl. Akad. Nauk SSSR* 117:2 (cit. on p. 31).

Tipping, Michael E. and Christopher M. Bishop (1999). "Probabilistic Principal Component Analysis". *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* 61:3, pp. 611–622. JSTOR: 2680726 (cit. on p. 136).

Trivedi, Shubhendu, Jialei Wang, Samory Kpotufe, and Gregory Shakhnarovich (23, 2014). "A Consistent Estimator of the Expected Gradient Outerproduct". In: *Proceedings of the*

Thirtieth Conference on Uncertainty in Artificial Intelligence. UAI'14. AUAI Press, Arlington, Virginia, USA, pp. 819–828 (cit. on p. 141).

Tsitouras, Ch. (1, 2011). "Runge–Kutta Pairs of Order 5(4) Satisfying Only the First Column Simplifying Assumption". *Computers & Mathematics with Applications* 62:2, pp. 770–775. doi: 10.1016/j.camwa.2011.06.002 (cit. on p. 131).

Tsybakov, A and V Zaiats (2009). *Introduction to Nonparametric Estimation* (cit. on p. 32).

Uno, Hajime, Tianxi Cai, Michael J. Pencina, Ralph B. D'Agostino, and L. J. Wei (10, 2011). "On the C-Statistics for Evaluating Overall Adequacy of Risk Prediction Procedures with Censored Survival Data". *Statistics in Medicine* 30:10, pp. 1105–1117. doi: 10.1002/sim.4154. pmid: 21484848 (cit. on p. 130).

Van Belle, Vanya, Kristiaan Pelckmans, Johan A. K. Suykens, and Sabine Van Huffel (2011). "Learning Transformation Models for Ranking and Survival Analysis". *Journal of Machine Learning Research* 12, pp. 819–862 (cit. on p. 60).

Van Belle, Vanya, Kristiaan Pelckmans, Johan A. K. Suykens, and Sabine Van Huffel (2007). "Support Vector Machines for Survival Analysis". *Proceedings of the Third International Conference on Computational Intelligence in Medicine and Healthcare* (cit. on p. 60).

Van der Laan, Mark J. and James M. Robins (2003). *Unified Methods for Censored Longitudinal Data and Causality*. 1st ed. Springer Series in Statistics. Springer New York, New York, NY. 1-694 (cit. on pp. 38, 60).

Van der Vaart, A. W. and Jon A. Wellner (1996). *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer Series in Statistics. Springer-Verlag, New York. doi: 10.1007/978-1-4757-2545-2 (cit. on pp. 162, 170).

Van Keilegom, Ingrid and Noël Veraverbeke (1, 1996). "Uniform Strong Convergence Results for the Conditional Kaplan-Meier Estimator and Its Quantiles". *Communications in Statistics - Theory and Methods* 25:10, pp. 2251–2265. doi: 10.1080/03610929608831836 (cit. on p. 37).

Vapnik, Vladimir (1998). *Statistical Learning Theory* (cit. on p. 30).

– (2000). *The Nature of Statistical Learning Theory*. 2nd ed. Information Science and Statistics. Springer-Verlag, New York. doi: 10.1007/978-1-4757-3264-1 (cit. on p. 30).

Vapnik, Vladimir and Alexey Chervonenkis (1974). *Теория Распознавания Образов: Статистические Проблеми Обучения*. Moscow (cit. on p. 66).

Vaswani, Ashish et al. (5, 2017). *Attention Is All You Need*. arXiv: 1706.03762 [cs]. url: http://arxiv.org/abs/1706.03762 (cit. on p. 131).

Wainwright, Martin J. (2019). *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge. doi: 10.1017/9781108627771 (cit. on pp. 29, 31, 32, 190).

Wand, M. P. and M. C. Jones (1, 1994). *Kernel Smoothing*. 60. Chapman & Hall, Boca Raton, FL, U.S. 224 pp. Google Books: GTOOi5yE008C (cit. on p. 43).

Wang, Yining, Simon Du, Sivaraman Balakrishnan, and Aarti Singh (31, 2018). "Stochastic Zeroth-Order Optimization in High Dimensions". In: *International Conference on Artificial Intelligence and Statistics*. International Conference on Artificial Intelligence and Statistics, pp. 1356–1365 (cit. on pp. 140, 153).

Wang, Yuxuan et al. (6, 2017). *Tacotron: Towards End-to-End Speech Synthesis*. arXiv: 1703.10135 [cs]. URL: http://arxiv.org/abs/1703.10135 (cit. on p. 116).

Wasserman, Larry (2004). *All of Statistics*. Springer Texts in Statistics. Springer New York, New York, NY. DOI: 10.1007/978-0-387-21736-9 (cit. on p. 32).

Wehenkel, Antoine and Gilles Louppe (12, 2021). *Graphical Normalizing Flows*. arXiv: 2006.02548 [cs, stat]. URL: http://arxiv.org/abs/2006.02548 (cit. on p. 120).

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, pp. 1871–1879. DOI: 10.1002/sim.4780111409 (cit. on p. 123).

Xia, Yingcun (2007). "A Constructive Approach to the Estimation of Dimension Reduction Directions". *Annals of Statistics* 35:6, pp. 2654–2690. DOI: 10.1214/009053607000000352 (cit. on p. 141).

Xia, Yingcun, Howell Tong, W. K. Li, and Li-Xing Zhu (2002). "An Adaptive Estimation of Dimension Reduction Space". *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* 64:3, pp. 363–410. JSTOR: 3088779 (cit. on p. 141).

Yang, Lei, Yixin Fang, Junhui Wang, and Yongzhao Shao (2017). "Variable Selection for Partially Linear Models via Learning Gradients". *Electronic Journal of Statistics* 11:2, pp. 2907–2930. DOI: 10.1214/17-EJS1300 (cit. on p. 137).

Ye, Gui-Bo and Xiaohui Xie (1, 2012). "Learning Sparse Gradients for Variable Selection and Dimension Reduction". *Machine Learning* 87:3, pp. 303–355. DOI: 10.1007/s10994-012-5284-9. arXiv: 1006.5060 (cit. on pp. 137, 141).

Zhou, Shanggang and Douglas A. Wolfe (2000). "On Derivative Estimation In Spline Regression". *Statistica Sinica* 10:1, pp. 93–108. JSTOR: 24306706 (cit. on p. 141).

Zonta, Tiago, Cristiano André da Costa, Rodrigo da Rosa Righi, Miromar José de Lima, Eduardo Silveira da Trindade, and Guann Pyng Li (1, 2020). "Predictive Maintenance in the Industry 4.0: A Systematic Literature Review". *Computers & Industrial Engineering* 150, p. 106889. DOI: 10.1016/j.cie.2020.106889 (cit. on pp. 2, 206).

Zou, Hui, Trevor Hastie, and Robert Tibshirani (2006). "Sparse Principal Component Analysis". *Journal of Computational and Graphical Statistics* 15:2, pp. 265–286. DOI: 10.1198/106186006X113430 (cit. on p. 135).

Добрушин, Р.Л (1953). "Обобщение уравнений Колмогорова для марковских процессов с конечным числом возможных состояний". *Матем. сб.* 33(75):3, pp. 567–596. DOI: 10.4213/rm1581 (cit. on p. 109).

Acronyms

ERM Empirical Risk Minimization

VC Vapnik-Chervonenkis

IPCW inverse probability of censoring weights

MAR missing at random

MCAR missing completely at random

MNAR missing not at random

LOO leave-one-out

k-**NN** *k*-nearest neighbours

TCGA The Cancer Genome Atlas

RNA ribonucleic acid

DNA deoxyribonucleic acid

AUC area-under-curve

SVR support-vector regression

RSF random survival forest

NPMLE nonparametric maximum likelihood estimation

EM expectation maximization

ODE ordinary differential equation

SABR "stochastic alpha, beta, rho"

SDE stochastic differential equation

PDE partial differential equation

GAN generative adversarial network

VAE variational autoencoder

AFT accelerated failure time

CRPS continuous ranked probability score

KL Kullback–Leibler

AD automatic differentiation

UDE universal differential equation

SDDE stochastic delay differential equation

LU lower-upper decomposition

IVP initial value problem

PCA principal components analysis

BS Brier score

BLL Binomial log-likelihood

IBS integrated Brier score

IBLL integrated Binomial log-likelihood

PPCA probabilistic principal component analysis

BRCA BReast CAncer gene

LLE locally linear embedding

UMAP uniform manifold approximation and projection for dimension reduction

SNE stochastic neighbor embedding

ELBO evidence lower bound

NF normalizing flow

CNF continuous normalizing flow

NLP natural language processing

LDA linear discriminant analysis

LASSO least absolute shrinkage and selection operator

EDR effective dimension reduction

EGOP expected gradient outerproduct

KD k-dimensional

PAC probably approximately correct

RWA risk weighted assets

BIS Bank for International Settlements

CDF cumulative distribution function

A-IRB advanced internal ratings-based

DNN deep neural network

ANN artificial neural network

FC fully connected

CART classification and regression tree

GGF gradient guided forest

RF random forest

ABS asset based security

VAR value-at-risk

MILP mixed integer linear programming

MCMC Markov chain Monte Carlo

CV cross-validation

GLM generalized linear model

GPU graphical processing unit

SIR Susceptible, Infectious, or Recovered

Glossary

λ Instantaneous hazard rate

Λ Cumulative hazard rate

$\hat{\Lambda}_n$ Kernel estimator of the cumulative hazard rate

X Covariates (random variable)

X_i Covariates of patient i (random variable)

β Parameter column vector of size d

$\mathcal{L}$ Loss of the form $\mathcal{L}(Y, f(X))$

$\mathbb{E}$ expectation

$\mathbb{P}$ Probability

$\mathbb{P}_n$ Empirical probability associated with the measure $\frac{1}{n}\sum_{i=1}^{n}\delta_{Y_i,X_i}$

$\mathbb{V}$ Variance

$\mathcal{R}$ Risk function $\mathcal{R}(f) = \operatorname{argmin} \mathbb{E}\left[\mathcal{L}(Y, f(X))\right]$

$\mathcal{R}_n$ Empirical risk function on $\mathcal{D}_n$, $\mathcal{R}_n(f) = \operatorname{argmin} \frac{1}{n}\sum_{i=1}^{n}\mathcal{L}(Y_i, f(X_i))$

$\widetilde{\mathcal{R}}_n$ IPCW empirical risk function on $\mathcal{D}_n$, $\mathcal{R}_n(f) = \operatorname{argmin} \frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i}{S_C(T_i|X_i)}\mathcal{L}(T_i, f(X_i))$

$\mathcal{R}_{\mathbb{K}}$ IPCW risk restricted to $\mathbb{K}$

$\widetilde{\mathcal{R}}_{n,\mathbb{K}}$ IPCW empirical risk restricted to $\mathbb{K}$

f Regression or predictive function $Y = f(X)$

f^* Minimizer of the risk $\operatorname{argmin}_f \mathcal{R}(f)$

$\bar{f}$ Minimizer of the restricted risk $\operatorname{argmin}_{f\in\mathcal{F}} \mathcal{R}(f)$

$\hat{f}_n$ Minimizer of the empirical restricted risk $\operatorname{argmin}_{f\in\mathcal{F}} \mathcal{R}_n(f)$

$\mathcal{F}$ Candidate class of functions of f the regression function, e.g. linear functions, trees etc.

Y Target variable, time-to-event in the survival setting

T Observed time such that $T = \min(Y, C)$

C Unobserved, nuisance, censoring variable

$p_C(\cdot \mid X)$ The conditional density of C, i.e. $p_C(t \mid X) = \mathbb{P}(C \in [t, t + \mathrm{d}t[\mid X)$

$p_Y(\cdot \mid X)$ The conditional density of Y, i.e. $p_Y(t \mid X) = \mathbb{P}(Y \in [t, t + \mathrm{d}t[\mid X)$

$S_C(\cdot \mid X)$ The conditional survival function of C, i.e. $S_C(t \mid X) = \int_t^\infty p_C(\mathrm{d}u \mid X)$

$S_Y(\cdot \mid X)$ The conditional survival function of Y, i.e. $S_Y(t \mid X) = \int_t^\infty p_Y(\mathrm{d}u \mid X)$

$\hat{S}_{A,n}(\cdot \mid X)$ Kernel estimator of the conditional survival function of the variable A

$\hat{S}_{C,m}(\cdot \mid X)$ Empirical estimator of $S_C(\cdot \mid X)$

$\hat{S}_{Y,m}(\cdot \mid X)$ Empirical estimator of $S_Y(\cdot \mid X)$

$\mathcal{E}$ Excess risk $\mathcal{E}(\hat{f}_n), \mathcal{F}) = \mathcal{R}(\hat{f}_n) - \mathcal{R}(\bar{f})$

$\mathfrak{R}$ Rademacher complexity

$\mathfrak{R}_n$ Empirical Rademacher complexity

$\mathcal{D}_n$ Set of training data $\{(Y_i, X_i, \delta_i)\}$

$\nu(\mathcal{F})$ vc dimension of $\mathcal{F}$

δ Censoring indicator $\delta = \mathbb{1}_{Y \leq C}$

$f * g$ Convolution of f and g

$S(t-)$ Left-limit of S at t

$H(u \mid X)$ Subsurvival $H(u \mid X) = \mathbb{P}(Y > u \mid X)$

$H_0(u \mid X)$ Subsurvival $H_0(u \mid X) = \mathbb{P}(Y > u, \delta = 0 \mid X)$

$\hat{H}_n(u \mid X)$ Kernel estimator of the subsurvival $H(u \mid X)$

$\hat{H}_{0,n}(u \mid X)$ Kernel estimator of the subsurvival $H_0(u \mid X)$

K Symmetric kernel function

K_h Scaled symmetric kernel function $K_h(x) = \frac{1}{h^d} K\left(\frac{x}{h}\right)$

$\mathbb{R}$ The reals.

$\mathbb{X}$ Space of X

$\mathbb{K}$ Compact subspace of $\mathbb{R} \times \mathbb{X}$

g Density of X

$\hat{g}_n$ Kernel estimator of the density of X

ϵ A Rademacher random variable in $\{-1, 1\}$

ε A supposedly small random variable or noise

$Z_n(\varphi)$ Random process of variations of the risk of φ (see chapter 2)

$\propto$ Proportional to

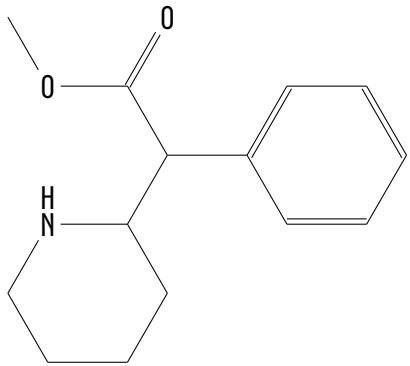
$M^\intercal$ Transposition of M

Index of Topics

V

Z

Index of Authors



non sic dormit, sed vigilat