Machine Learning Approaches for the Prediction of Credit Risk

Thesis Defense

BNP PARIBAS

Guillaume Ausset

Télécom Paris, BNP Paribas

ECOLE
DOCTORALE
DE MATHEMATIQUES
HADAMARD

Agenda for the Day

- **Credit Risk: Predict, not Estimate**

  ⇒ Individualized/Conditional models.

  ⇒ Adapt the ERM approach.

- **Empirical Risk Minimization with Reweighting**

  ⇒ Correct for censoring by the inverse probability of censoring, derive generalization bounds.

- **Survival Normalizing Flows**

  ⇒ Flexible generative model for survival.

  ⇒ Tractable likelihood.

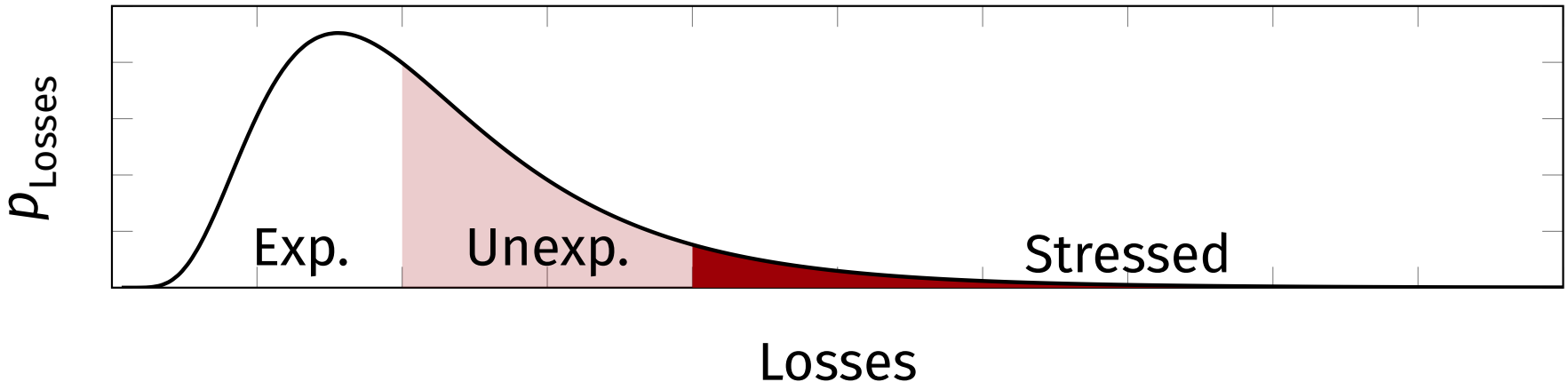
- **Gradient Based Approach to Dimension Reduction**

  ⇒ Estimate the gradient, select variables.

Credit Risk:
Predict, not Estimate

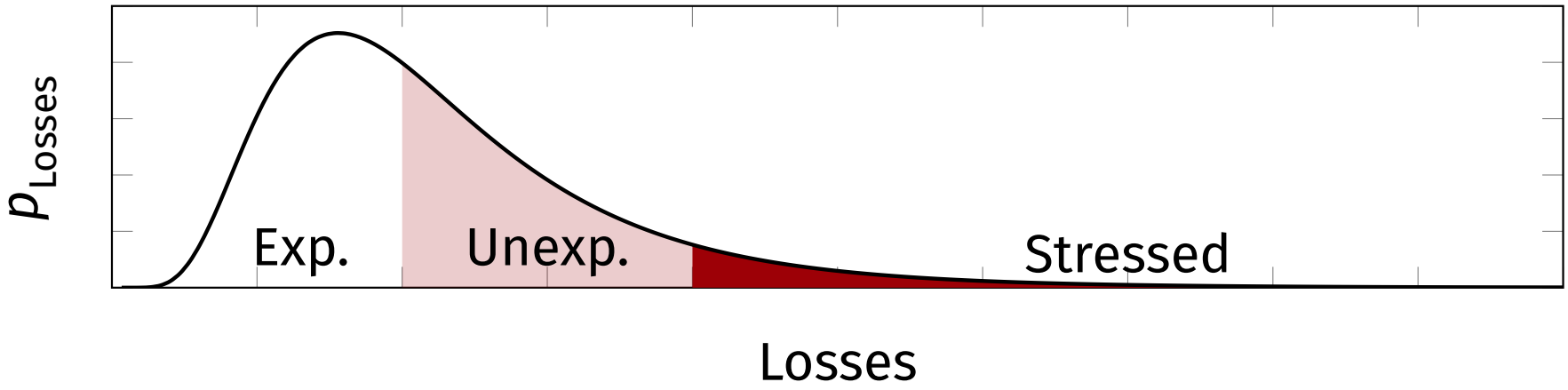
Credit Risk at BNP Paribas

BNP Paribas manages a porfolio of credits.
⇒ accumulation of risky assets.
⇒ Regulator requires Capital to survive losses.
Portfolio Management's role is measuring and reducing this risk.



Credit Risk at BNP Paribas

BNP Paribas manages a porfolio of credits.
⇒ accumulation of risky assets.
⇒ Regulator requires Capital to survive losses.
Portfolio Management's role is measuring and reducing this risk.



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

(Embrechts, Klüppelberg, and Mikosch 1997; Basel Committee 2019)

Approaches for the Estimation of p

Structural Risk Models (Merton 1974)

⇒ Rigid structure. Calibration on market data needed.

Market Approach (using CDS Bielecki and Rutkowski 2004)

⇒ CDS inexistent or illiquid.

Aggregated Portfolio Models (Lopez and Saidenberg 1999)

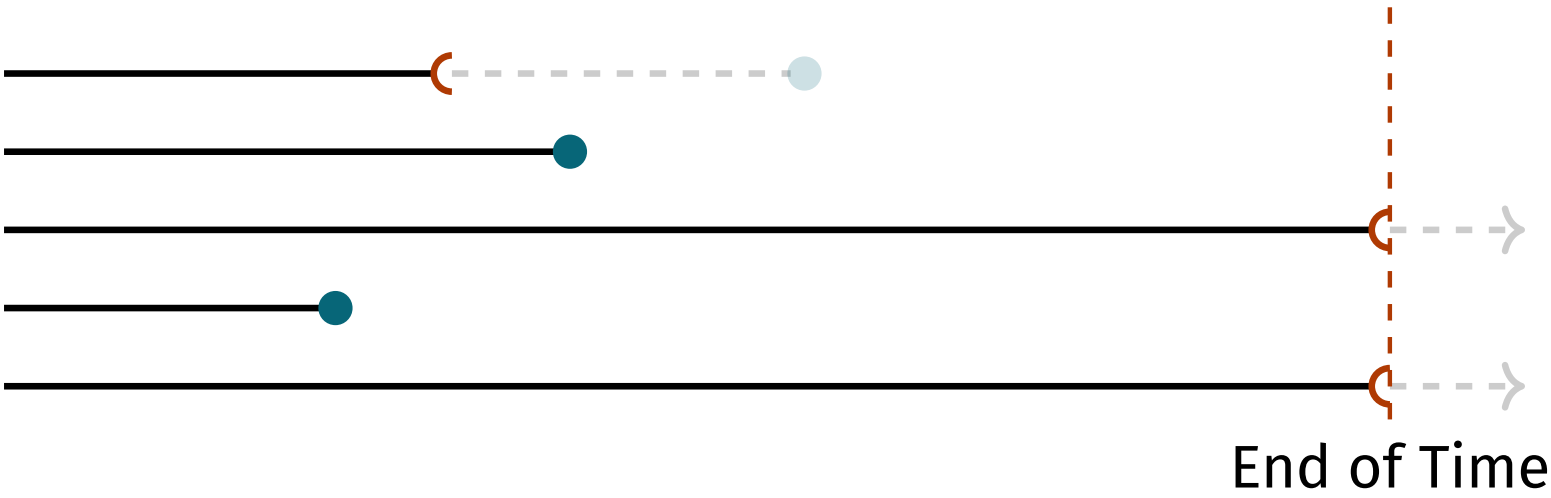
⇒ Not Individual / Conditional.

Individual Models (Avesani, Liu, and Mirestean 2006)

⇒ Usually classification problems. Arbitrary discretization and severe class imbalance.

Censored Observations

Not a classification problem, a time-to-event problem.
Only observe (T, δ, X), not (Y, X).

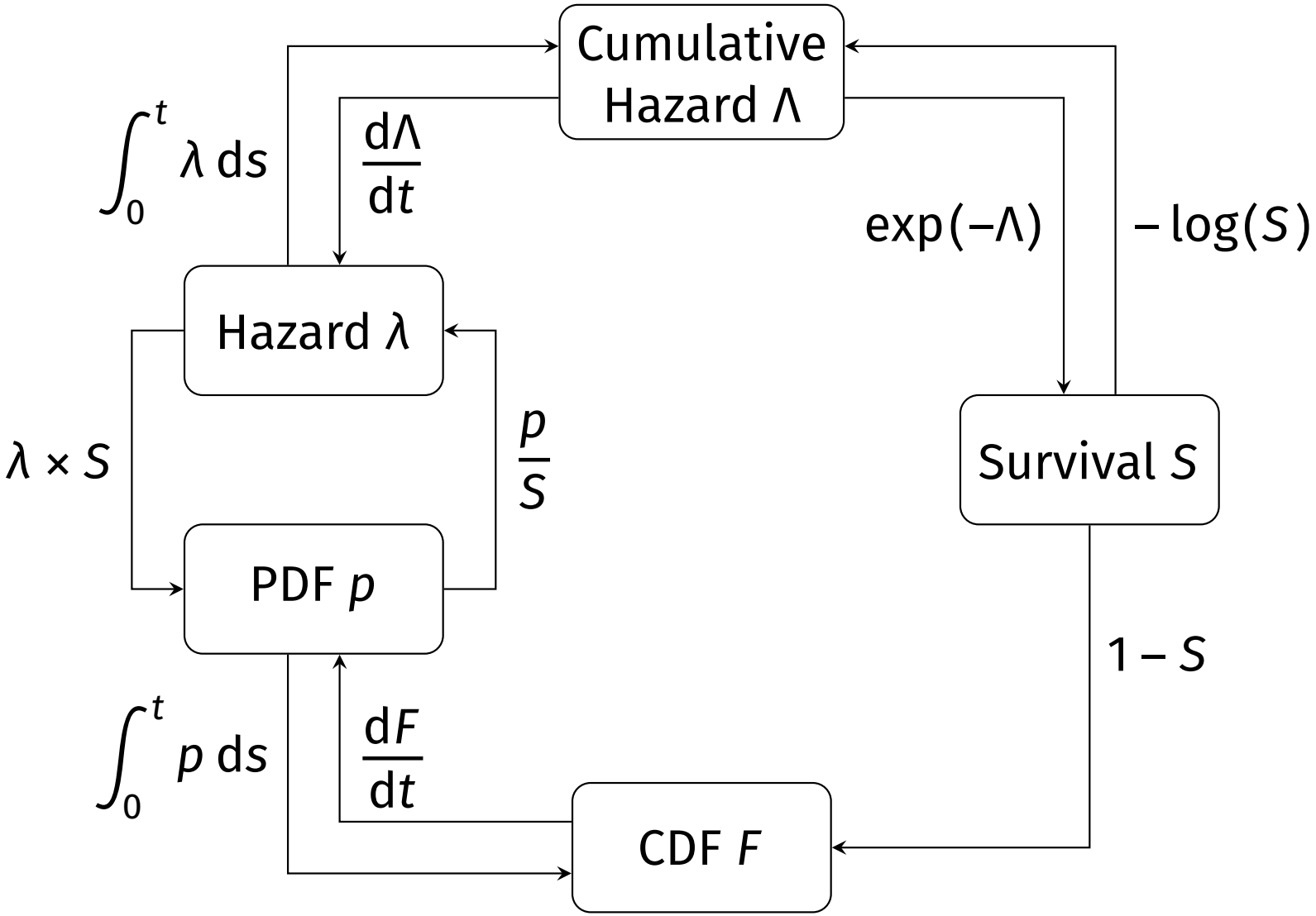


$$T = \min(Y, C)$$

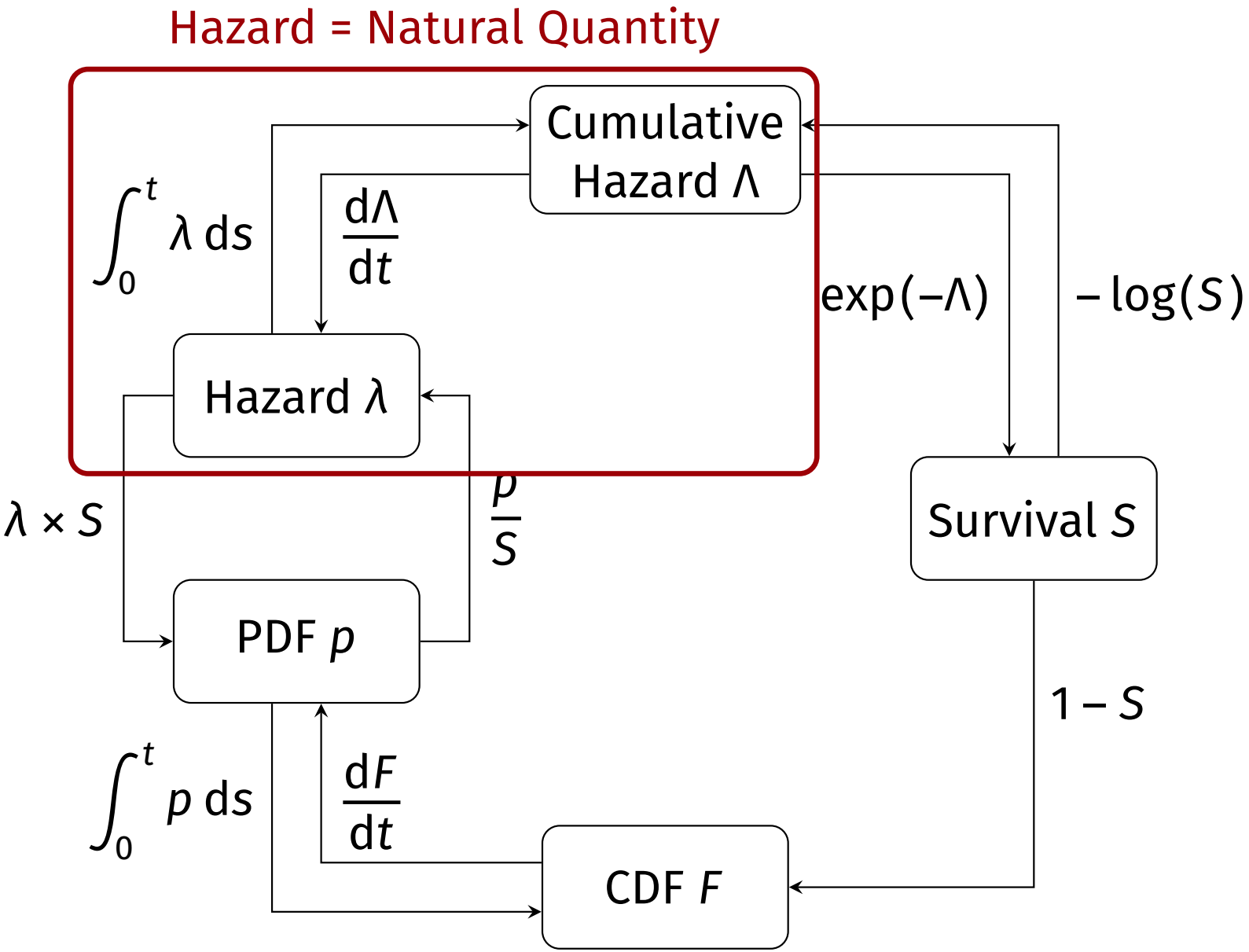
$$\delta = \mathbb{1}_{Y \leq C} = \begin{cases} 1 & \text{if } Y \leq C \\ 0 & \text{otherwise.} \end{cases}$$

Also: left-censoring, left/right truncations.

What to estimate?



What to estimate?



Non-Parametric Approach

Leads to the Kaplan-Meier estimator of S and Nelson-Aalen estimator of Λ:

$$\hat{S}_n(t) = \prod_{i=1}^{n} \left(1 - \frac{\delta_{[i]}}{n-i+1}\right)^{\mathbb{1}_{T_{[i]} \leq t}}$$

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

$T_{[i]}$ ordered time (Andersen et al. 1993; Fleming and Harrington 1991).

Non-Parametric Approach

Leads to the Kaplan-Meier estimator of S and Nelson-Aalen estimator of Λ:

$$\hat{S}_n(t) = \prod_{i=1}^{n} \left(1 - \frac{\delta_{[i]}}{n-i+1}\right)^{\mathbb{1}_{T_{[i]} \leq t}}$$

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

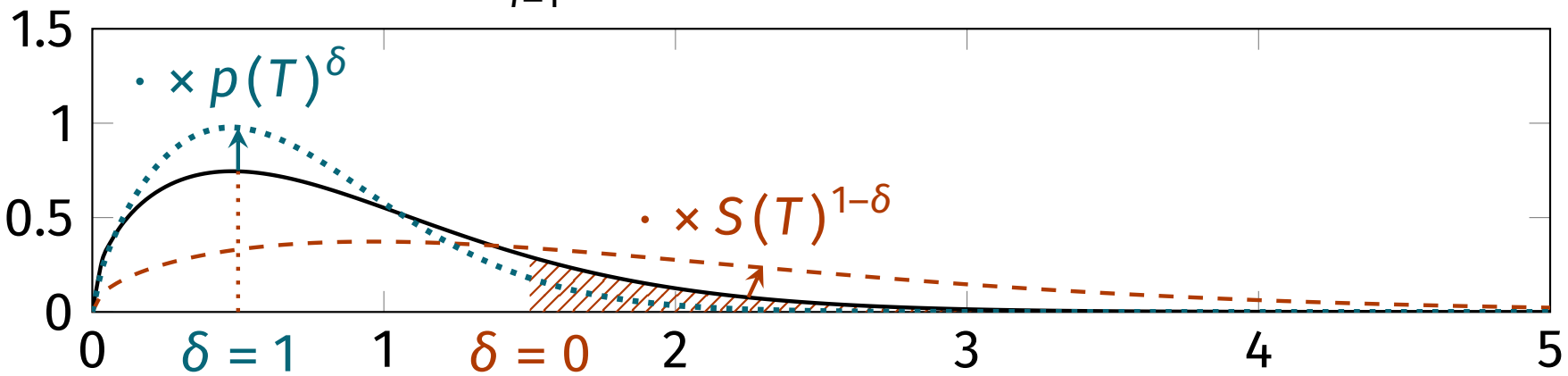
$T_{[i]}$ ordered time (Andersen et al. 1993; Fleming and Harrington 1991). Can also be conditioned with kernel K:

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

(van Keilegom 1998)

Parametric Approach

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



Can learn parametric estimators, e.g. (Cox 1972):

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

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

Prediction, not estimation

Not interested in estimating the distribution, only to predict $Y = f(X)$ the event.

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

One approach, the risk minimization framework:

$$f^* = \operatorname*{argmin}_f \mathbb{E}[L(Y, f(X))]$$

Minimizes a loss L, e.g.

$$f(x) = \mathbb{E}[Y \mid X = x] = \operatorname{argmin} \mathbb{E}[(Y - f(X)^2)].$$

Prediction, not estimation

Not interested in estimating the distribution, only to predict $Y = f(X)$ the event.

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

One approach, the risk minimization framework:

$$f^* = \underset{f}{\operatorname{argmin}}\, \mathbb{E}[L(Y, f(X))]$$

Minimizes a loss L, e.g.

$$f(x) = \mathbb{E}[Y \mid X = x] = \operatorname{argmin} \mathbb{E}[(Y - f(X)^2)].$$

Y is unknown.

Empirical Risk Minimization with Reweighting

Inverse Probability Reweighting

True time-to-event is unknown.

$$R(f) = \mathbb{E}\Big[(Y - f(X))^2\Big] \xrightarrow{\text{Conditioning}} R(f) = \mathbb{E}\left[\frac{\delta(T - f(X))^2}{S_C(T- \mid X)}\right].$$

$$R_n(f) = \sum_{i=1}^{n}\big(Y_i - f(X_i)\big)^2 \longrightarrow R_n(f) = \frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i(T_i - f(X_i))^2}{S_C(T_i- \mid X_i)}.$$

Inverse Probability Reweighting

True time-to-event is unknown.

$$R(f) = \mathbb{E}\Big[(Y - f(X))^2\Big] \xrightarrow{\text{Conditioning}} R(f) = \mathbb{E}\left[\frac{\delta(T - f(X))^2}{S_C(T- \mid X)}\right].$$

$$R_n(f) = \sum_{i=1}^{n}\big(Y_i - f(X_i)\big)^2 \Longrightarrow R_n(f) = \frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i(T_i - f(X_i))^2}{S_C(T_i- \mid X_i)}.$$

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

Inverse Probability Reweighting

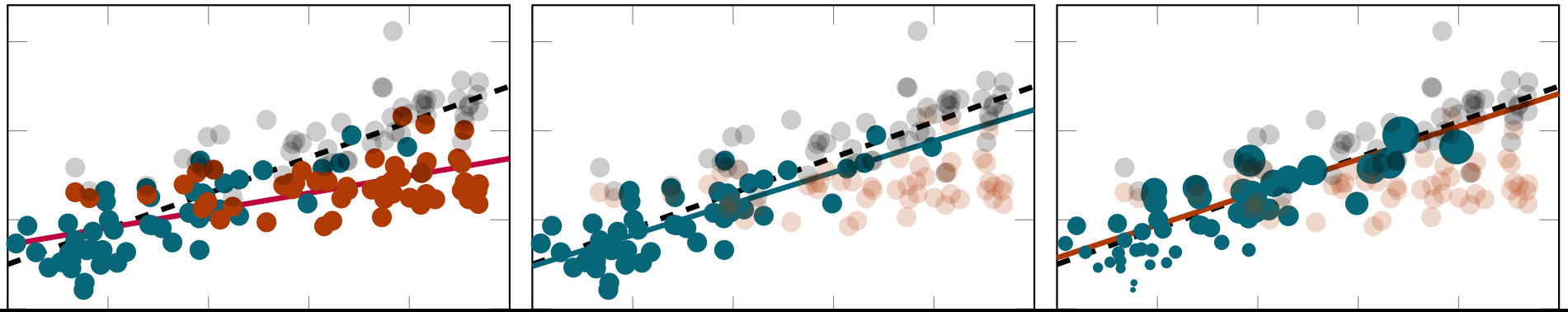
True time-to-event is unknown.

$$R(f) = \mathbb{E}\left[(Y - f(X))^2\right] \xrightarrow{\text{Conditioning}} R(f) = \mathbb{E}\left[\frac{\delta(T - f(X))^2}{S_C(T- \mid X)}\right].$$

$$R_n(f) = \sum_{i=1}^{n}\left(Y_i - f(X_i)\right)^2 \Longrightarrow R_n(f) = \frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i(T_i - f(X_i))^2}{S_C(T_i- \mid X_i)}.$$

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

Learn on Censored | Learn on Observed | Learn on Reweighted



Prior Work

Can be seen as a weighted ERM problem:

$$\sum_{i=1}^{n} \omega_i \big(T_i - f(X_i)\big)^2 \qquad \text{with} \qquad \omega_i = \frac{\delta_i}{S_C(T_i - \mid X_i)}.$$

Results assume that ω_i is known (Cortes, Mansour, and Mohri 2010). IPCW approach studied asymptotically and with strong $Y \perp\!\!\!\perp C$ hypothesis.

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

Stute and J.-L. Wang 1993; Stute 2003, 1995b,a, 1993b,a, 1996 and non-asymptotic but non-uniform results Dabrowska 1989, 1988, 1986.

Kernel Estimators of the (sub-) Survival

Use a kernel conditional estimator in the style of Kaplan-Meier:

$$\hat{\Lambda}_{C,n}(u \mid x) = -\int_0^u \frac{\sum_{i=1}^n K_h(x - X_i)\mathbb{1}_{T_i>\mathrm{d}s,\ \delta_i=0}}{\sum_{i=1}^n K_h(x - X_i)\mathbb{1}_{T_i>s-}},$$

$$\hat{S}_{C,n}(u \mid x) = \prod_{s\leq u}\Big(1 - \mathrm{d}\hat{\Lambda}_{C,n}(s \mid x)\Big),$$

Kernel Estimators of the (sub-) Survival

Use a kernel conditional estimator in the style of Kaplan-Meier:

$$\hat{\Lambda}_{C,n}(u \mid x) = -\int_0^u \frac{\sum_{i=1}^n K_h(x - X_i)\mathbb{1}_{T_i > \mathrm{d}s,\ \delta_i = 0}}{\sum_{i=1}^n K_h(x - X_i)\mathbb{1}_{T_i > s-}},$$

$$\hat{S}_{C,n}(u \mid x) = \prod_{s \leq u}\Big(1 - \mathrm{d}\hat{\Lambda}_{C,n}(s \mid x)\Big),$$

Plug it in the loss:

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

Problem: Not an i.i.d. sum.

Uniform control of the excess risk

Ausset, Clémençon, and Portier 2021a, Theorem 2.9

Suppose that $Y \perp\!\!\!\perp C \mid X$ and F is a bounded vc class. There exist constants h_0, M_1, M_2 and M_3, such that, for all $n \geq 2$ and $\varepsilon \in (0,1)$, the event

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

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

Same rates as without censoring !

Faster rates than expected.

Uniform control of the survival function

Ausset, Clémençon, and Portier 2021a, Theorem 2.6

Suppose that $Y \perp\!\!\!\perp C \mid X$ and $x \mapsto H(u \mid x), x \mapsto H_0(u \mid x)$ and $x \mapsto g(x)$ are smooth. Then, there exist constants $M_1 > 0$, $M_2 > 0$ and $h_0 > 0$ 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 $nh^d \geq M_2 \left|\log(h^{d/2}\varepsilon)\right|$.

Proof idea: Linearize

We want to control the random process

$$Z_n(\varphi) = \underbrace{\left(\frac{1}{n} \sum_{i=1}^{n} \frac{\delta_i \varphi(T_i, X_i)}{\hat{S}_{C,n}^{(i)}(T_i - \mid X_i)} \right)}_{\text{Not an i.i.d. sum.}} - \mathbb{E}[\varphi(Y, X)], \ \varphi \in \Phi,$$

Linearize by Taylor extensions:

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

Where

- $L_n(\varphi)$, a centered i.i.d. sum.
- $M_n(\varphi)$, sum of centered i.i.d. sum, bias term, U-statistics and residual.
- $R_n(\varphi)$, a residual.

Proof idea: Control Independent Parts

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

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

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

Proof idea: Control Independent Parts

$$L_n(\varphi) = \underbrace{\frac{1}{n}\sum_{i=1}^{n}\left(\delta_i \frac{\varphi(T_i, X_i)}{S_C(T_i \mid X_i)} - \mathbb{E}\left[\delta \frac{\varphi(T, X)}{S_C(T \mid X)}\right]\right)}_{\text{i.i.d. centered sum, } 1/\sqrt{n}},$$

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

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

Proof idea: Control Independent Parts

$$L_n(\varphi) = \underbrace{\frac{1}{n}\sum_{i=1}^{n}\left(\delta_i \frac{\varphi(T_i, X_i)}{S_C(T_i \mid X_i)} - \mathbb{E}\left[\delta \frac{\varphi(T, X)}{S_C(T \mid X)}\right]\right)}_{\text{i.i.d. centered sum, } 1/\sqrt{n}},$$

$$M_n(\varphi) = \underbrace{-\frac{1}{n}\sum_{i=1}^{n}\delta_i \varphi(T_i, X_i)\frac{\hat{a}_n^{(i)}(T_i \mid X_i)}{S_C(T_i \mid X_i)}}_{\text{Decompose in U-statistic and residual.}},$$

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

Proof idea: Control Independent Parts

$$L_n(\varphi) = \underbrace{\frac{1}{n}\sum_{i=1}^{n}\left(\delta_i \frac{\varphi(T_i, X_i)}{S_C(T_i \mid X_i)} - \mathbb{E}\left[\delta \frac{\varphi(T, X)}{S_C(T \mid X)}\right]\right)}_{\text{i.i.d. centered sum, } 1/\sqrt{n}},$$

$$M_n(\varphi) = \underbrace{-\frac{1}{n}\sum_{i=1}^{n}\delta_i \varphi(T_i, X_i)\frac{\hat{a}_n^{(i)}(T_i \mid X_i)}{S_C(T_i \mid X_i)}}_{\text{Decompose in U-statistic and residual.}},$$

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

Proof idea: Tools

1. Control deterministic parts by integration results.
   ⇒ use e.g. Delyon and Portier 2016
2. Control vc class of functionals involved.
   ⇒ use Giné and Guillou 2001; Giné and Sang 2010.
3. Control variations of M / U-statistics.
   ⇒ use Major 2006.
4. Bound the residuals.

Experimental Results

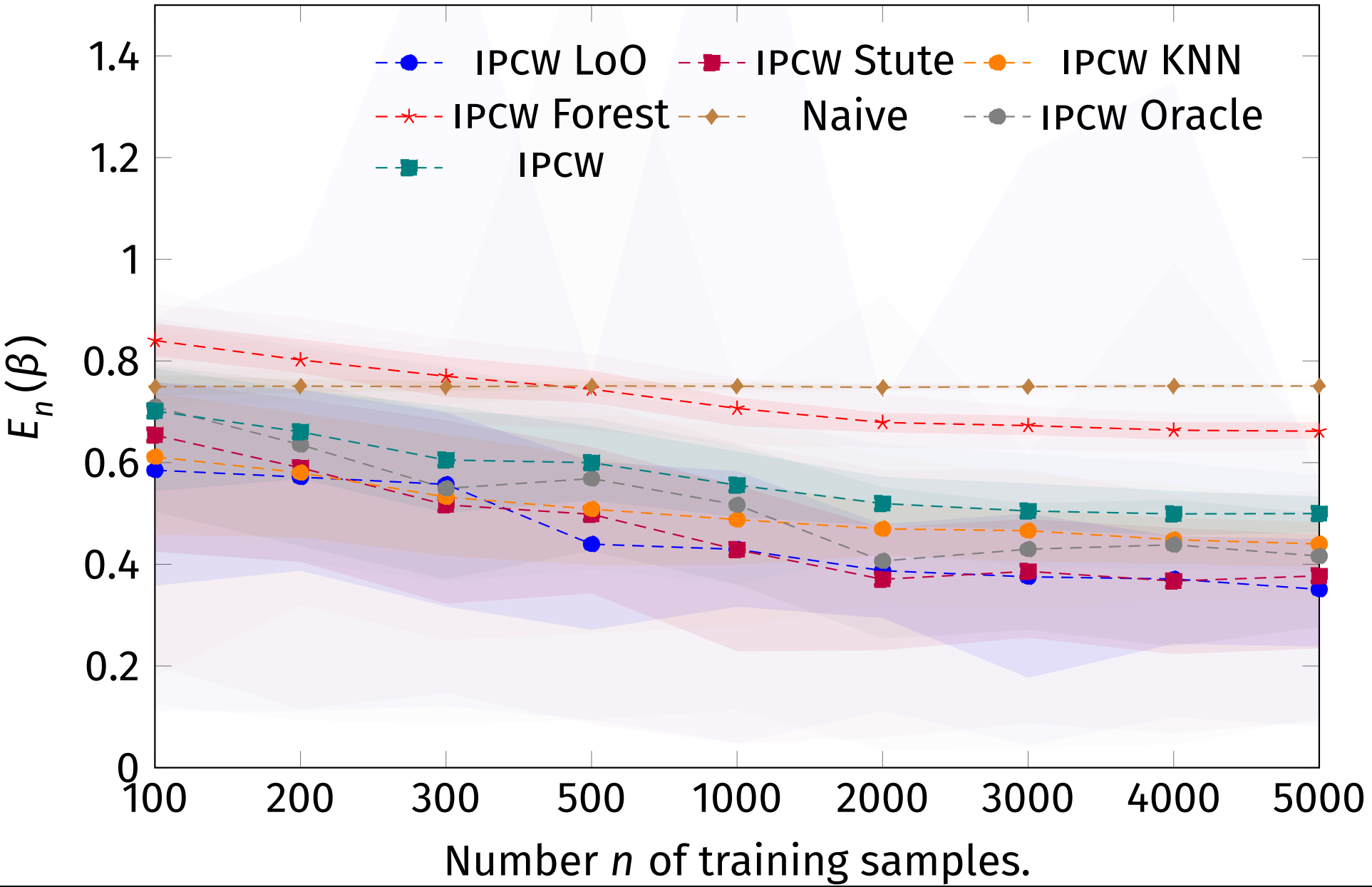
$$\text{Kernel} \quad \sum_{i=1}^{n} \delta_i \frac{\big(T_i - f(X_i)\big)^2}{\hat{S}_{C,n}^{\texttt{Kern}}(T_i \mid X_i)} \qquad \text{LOO} \quad \sum_{i=1}^{n} \delta_i \frac{\big(T_i - f(X_i)\big)^2}{\hat{S}_{C,n}^{(i)\texttt{Kern}}(T_i \mid X_i)}$$

$$\text{Forest} \quad \sum_{i=1}^{n} \delta_i \frac{\big(T_i - f(X_i)\big)^2}{\hat{S}_{C,n}^{\texttt{RF}}(T_i \mid X_i)} \qquad \text{Stute} \quad \sum_{i=1}^{n} \delta_i \frac{\big(T_i - f(X_i)\big)^2}{\hat{S}_{C,n}^{\texttt{KM}}(T_i)}$$

$$\text{KNN} \quad \sum_{i=1}^{n} \delta_i \frac{\big(T_i - f(X_i)\big)^2}{\hat{S}_{C,n}^{(i)\texttt{KNN}}(T_i \mid X_i)} \qquad \text{Oracle} \quad \sum_{i=1}^{n} \delta_i \frac{\big(T_i - f(X_i)\big)^2}{S_C(T_i \mid X_i)},$$

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

Experimental Results



Experimental Results: Synthetic Dataset

	Method	$p = 3/4$	$p = 1/2$	$p = 1/4$
`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

Survival Normalizing Flows

More Flexible Estimators of the Survival

- IPCW depends on $\hat{S}_{C,n}$.

  $\Rightarrow$ Need better estimators.
- Complex unstructured (e.g. text) in finance.

  $\Rightarrow$ Neural Estimators.
- Simulations required in finance.

  $\Rightarrow$ Generative models.
- Bayesian modelling for low data regime.

  $\Rightarrow$ Tractable likelihood needed.

The Change of Variable Theorem.

Model the survival Y as the mapping of a latent variable:

$$Y = m(Z, X).$$

The Change of Variable Theorem.

Model the survival Y as the mapping of a latent variable:

$$Y = m(Z, X).$$

If m, C^1-diffeomorphism, change of variable theorem (Rezende and Mohamed 2015):

$$\log p_Y(t) = \log p_Z(z) - \log \left| \det \frac{\partial m_\theta}{\partial z} \right|.$$

Problem: Computing determinant expensive ($O(n^3)$)

Solution: Triangular Jacobian

(Kingma and Dhariwal 2018; Papamakarios, Pavlakou, and Murray 2018; Wehenkel and Louppe 2021; Durkan et al. 2019)

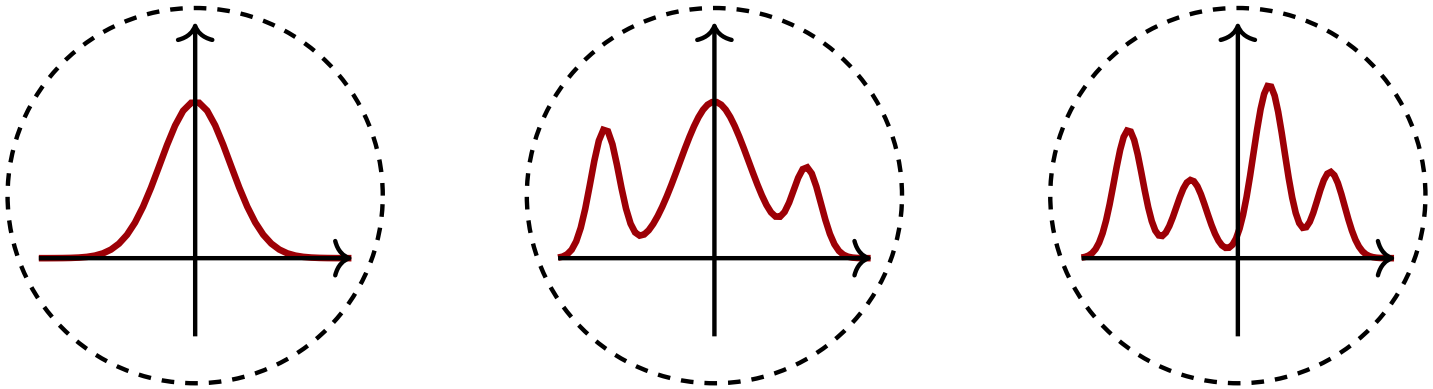
Discrete Normalizing Flows

New Problem: m too simple.
New New Solution: Compose many m.

$$Y = m_{\theta,K} \circ \ldots \circ m_{\theta,0}(Z),$$

$$\log p_Y(t) = \log p_Z(z) - \sum_{i=0}^{K} \log \left| \det \frac{\partial m_{\theta,i}}{\partial z_i} \right|.$$

$$z = z_0 \xrightarrow{m_{\theta,1}} z_1 \dashrightarrow z_i \xrightarrow{m_{\theta,i+1}} z_{i+1} \dashrightarrow z_k = y$$



$z_0 \sim p_0(z_0)$ $z_i \sim p_i(z_i)$ $z_k \sim p_k(z_k)$

Continuous Normalizing Flows

Other solution: continuous change of variable theorem:

$$\begin{cases} z_{i+1} = m_{\theta,i}(z_i) \\ \dfrac{\log p(z_{i+1}) - \log p(z_i)}{i+1-i} = -\log \left| \dfrac{\partial m_{\theta,i}}{\partial z} \right| \end{cases} \Rightarrow \begin{cases} \dfrac{\partial \mathbf{z}_\theta}{\partial t} = m_\theta(\mathbf{z}_\theta(t), t) \\ \dfrac{\partial \log p(\mathbf{z}_\theta(t))}{\partial t} = -\operatorname{tr} \dfrac{\partial m_\theta}{\partial \mathbf{z}} \end{cases}$$

Trace costs $O(n)$. Invertibility not needed explicitly.

$$\frac{\partial}{\partial t} \begin{bmatrix} \mathbf{z}_\theta(t) \\ \log p(y) - \log p(\mathbf{z}_\theta(t)) \end{bmatrix} = \begin{bmatrix} m_\theta(\mathbf{z}_\theta(t), t) \\ -\operatorname{tr} \dfrac{\partial m_\theta}{\partial \mathbf{z}} \end{bmatrix}$$

$$\begin{bmatrix} \mathbf{z}_\theta(1) \\ \log p(y) - \log p(\mathbf{z}_\theta(1)) \end{bmatrix} = \begin{bmatrix} y \\ 0 \end{bmatrix}.$$

Solve ODE. Reverse dynamics to invert. (Chen et al. 2018)

Conditional Survival Normalizing Flows

Ausset, Ciffreo, et al. 2021

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

- Compute $p(T_i \mid X, \theta)$ by ODE.
- Compute $S(T_i \mid X, \theta)$ on latent: $S_Y(Y_i \mid X) = S_Z\left(M_\theta^{-1}(Y_i, X)\right)$.
- Differentiate L using method of adjoints (Cao et al. 2003; Rackauckas, Ma, Dixit, et al. 2018).

Conditional Survival Normalizing Flows

Ausset, Ciffreo, et al. 2021

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

- Compute $p(T_i \mid X, \theta)$ by ODE.
- Compute $S(T_i \mid X, \theta)$ on latent: $S_Y(Y_i \mid X) = S_Z\left(M_\theta^{-1}(Y_i, X)\right)$.
- Differentiate L using method of adjoints (Cao et al. 2003; Rackauckas, Ma, Dixit, et al. 2018).

Introduce conditional version of m:

$$m_\theta(\mathbf{z}_\theta(t,x),t,x) = \sum_i \underbrace{\pi_{\theta,i}(x)}_{\text{Precompute.}} \underbrace{\sigma_{\theta,i}(t) m_{\theta,i}(\mathbf{z}_\theta(t,x))}_{\text{Adaptive solver.}}.$$

$\Rightarrow$ Can be trained on a GPU.

Survival Flows on Real Data.

Compared to SoTA on the Concordance

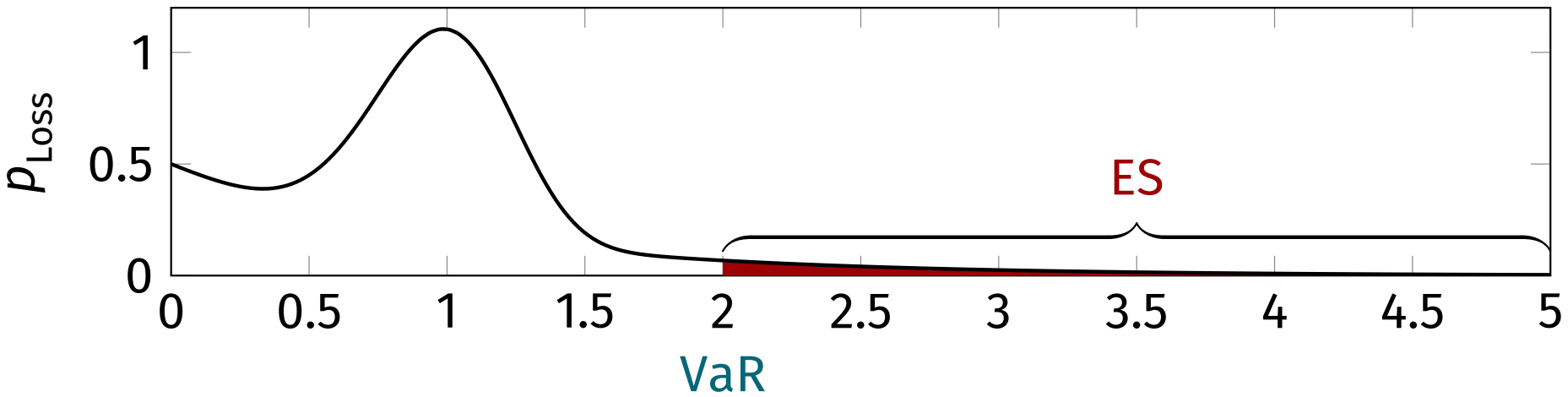
$$C(S) = \frac{\sum_i \sum_j \delta_j \mathbb{1}_{S(X_j) > S(X_i)} \mathbb{1}_{T_j \leq T_i}}{\sum_i \sum_j \delta_j \mathbb{1}_{T_j \leq T_i}},$$

	Concordance			
Method	Support	WHAS	RGBSG	Metabric
This work	0.61678	0.86059	**0.68464**	**0.64879**
DeepSurv	**0.61831**	0.86262	0.66840	0.64337
RSF	0.61302	**0.89362**	0.65119	0.62433
Cox PH	0.58287	0.81762	0.65775	0.63062

Portfolio Optimization

$$\sum_{i=1}^{n} \omega_i c_i, \text{ with } L(\omega) = \sum_{i=1}^{K} \omega_i (l_i \mathbb{1}_{Y_i \leq d_i} - c_i).$$

Criterion: $\text{ES}_\alpha(\omega) = \int_{-\infty}^{\alpha} F^{-1}_{L(\omega)}(\gamma)\, \mathrm{d}\gamma$

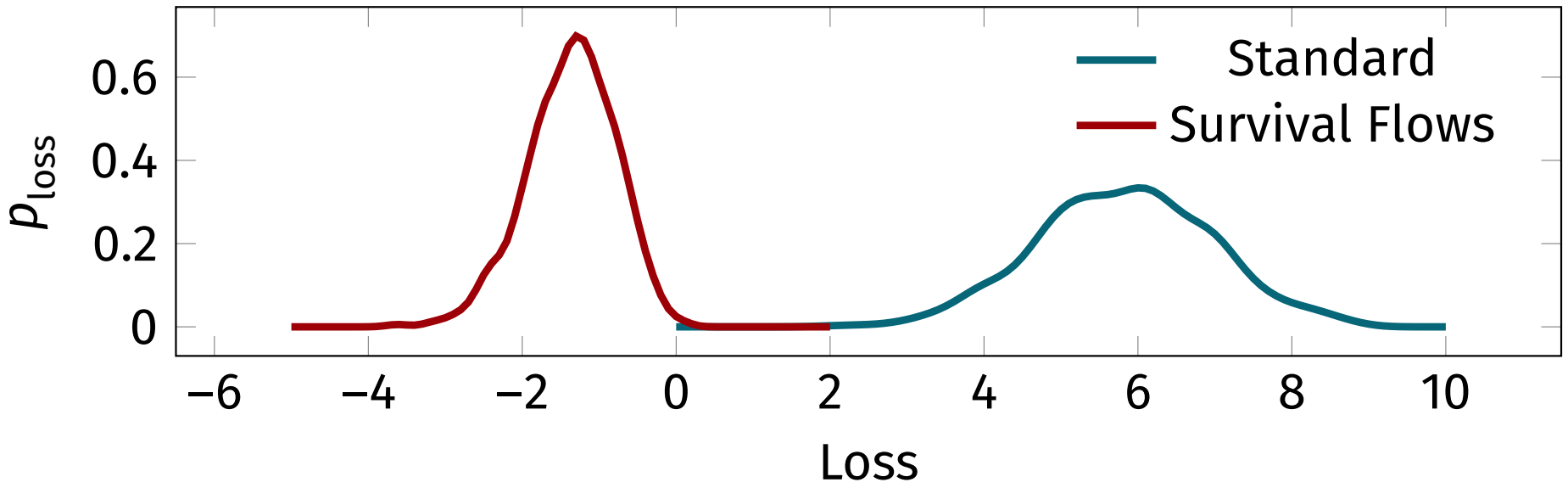


(Artzner et al. 1999; Embrechts, Klüppelberg, and Mikosch 1997)

From Expected Shortfall to LP

$\text{ES}_\alpha(\omega)$ solution of LP $\Rightarrow$ Portfolio Optimization also LP (Rockafellar and Uryasev 2002).

$$\begin{aligned} \underset{\omega,\beta}{\text{argmin}} \quad & \beta + \frac{1}{1-\alpha} \int [L(\omega) - \beta]^+ p_L(y)\, \mathrm{d}y \\ \text{s.t.} \quad & 0 \leq \omega \leq 1 \\ & \omega^\intercal c = P. \end{aligned}$$



Gradient Based Approach to Dimension Reduction

Supervised vs Unsupervised

- Financial (e.g. text) highly dimensional.
- Detrimental to predictive / computational performance.

Need to reduce dimension.

Unsupervised Use only X.

⇒ PCA, VAE, SNE ...

Supervised Use Y & X.

⇒ LDA, Representation Learning, Effective Dimension Reduction (EDR) ≈ Gradient directions.

Gradient Based EDR

Single-index model:

$$Y_i = f(X_i) = g(AX_i) + \varepsilon_i,$$

Space spanned by A also spanned by gradient.

Retrieve principal components of

$$M = \frac{1}{n} \sum_{i=1}^{n} \nabla f(X_i) \nabla^\intercal f(X_i).$$

(Hristache, Juditsky, and Spokoiny 1998; Hristache, Juditsky, Polzehl, et al. 2001; Bücher, El Ghouch, and Van Keilegom 2014)

Can be extended to manifolds. (Mukherjee, Wu, and Zhou 2010; Aswani, Bickel, and Tomlin 2011)

Local Linear Estimator

$$f(z) = f(x) + \nabla f(x)^{\intercal}(z - x) + o(\|z - x\|).$$

For any point X_i close to x,

$$f(X_i) \approx f(x) + \beta^{\intercal}\left(X_i - x\right)$$

Idea: Solve ERM on k-NN neighbourhood $\hat{\imath}_k(x)$ of x.

$$\underset{(\alpha,\beta)\in\mathbb{R}^{d+1}}{\operatorname{argmin}} \sum_{i\in\hat{\imath}_k(x)} \left(Y_i - \alpha - \beta^{\intercal}(X_i - x)\right)^2.$$

Add LASSO to retrieve sparsity:

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

(Fan 1992, 1993)

Compared to existing approaches

Results exist on local-linear estimators with kernel averaging. (Fan 1993; Fan and Gijbels 1996; Dalalyan, Juditsky, and Spokoiny 2008).

We add:

- k-NN neighbourhood.
  - ⇒ Use results from Jiang 2019.
  - ⇒ Control the random radius of k-NN the balls.
- LASSO to retrieve sparsity, improve bounds.

Compared to existing approaches

Results exist on local-linear estimators with kernel averaging. (Fan 1993; Fan and Gijbels 1996; Dalalyan, Juditsky, and Spokoiny 2008).

We add:

- k-NN neighbourhood.
  - ⇒ Use results from Jiang 2019.
  - ⇒ Control the random radius of k-NN the balls.
- LASSO to retrieve sparsity, improve bounds.

Only mild assumptions:

- Boundedness assumptions on density.
- Sub-gaussianity of residuals $(Y - f(x))$.
- Lipschitzianity of residuals.

Bounds on the Gradient Estimation

Ausset, Clémençon, and Portier 2021b, Theorem 4.1

Suppose that the previous mild assumptions are fulfilled. Let $n \geq 1$ and $k \geq 1$ such that $\overline{\tau}_k \leq \tau_0$.

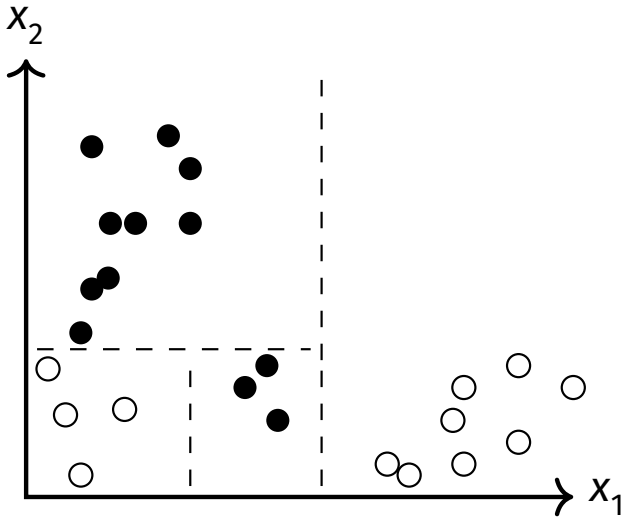
$$\lambda = \overline{\tau}_k \left(\sqrt{2\sigma^2 \frac{\log(16d/\delta)}{k}} + L_2 \overline{\tau}_k^2 \right).$$

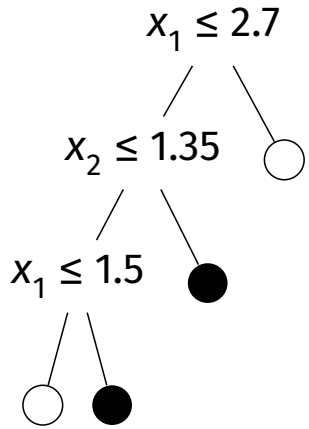
Then, we have with probability larger than $1 - \delta$,

$$\left\| \tilde{\beta}_k(x) - \beta(x) \right\|_2 \leq 24^2 \sqrt{|S_x|} \left(\overline{\tau}_k^{-1} \sqrt{\frac{2\sigma^2 \log(16d/\delta)}{k}} + L_2 \overline{\tau}_k \right),$$

as soon as $C_1 |S_x| \log\left(\frac{dn}{\delta}\right) \leq k \leq C_2 n$, and $\overline{\tau}_k^2 \leq \frac{b_f^2}{C_3 |S_x| L^2} \wedge \tau_0^2$, where C_1, C_2 and C_3 are universal constants.

Gradient Guided Forests





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: $\omega \leftarrow \sum_{i \in \text{Node}} |\nabla r(X_i)|$
3: $K \leftarrow$ `sample` $\sqrt{d}$ `dimensions in` $\{1, ..., d\}$ `with probabilities` $\propto \omega$
4: $k, c \leftarrow$ `best threshold` c `and dimension` k
5: **return** k, c

Exploit locality of variable selection.

Gradient Guided Forests

	Description		Loss	
Dataset	n	d	RF	GGF
Wisconsin	569	30	0.0352	**0.0345**
Heart	303	13	0.128	**0.124**
Diamonds	53940	23	680033	**664265**
Gasoline	60	401	0.678	**0.512**
SDSS	10000	8	$0.872 \cdot 10^{-3}$	$\mathbf{0.776 \cdot 10^{-3}}$

Guiding cuts on relevant variables improves performance.

Blackbox Optimization

$$x_{n+1} = x_n - \gamma \underbrace{\nabla f(x_n)}_{\text{Unknown}}$$

Require: x_0: initial guess, f: function $\mathbb{R}^d \mapsto \mathbb{R}$, M: budget

1: $X \leftarrow X_1, ..., X_M$ `with` $X_i \sim N(x_0, \varepsilon \times I_d)$
2: $Y \leftarrow f(X) \coloneqq f(X_1), ..., f(X_M)$
3: **while** `not StoppingCondition` **do**
4: $r, \delta \leftarrow$ `estimated gradient at` x `w.r.t` X, Y `using` eq. (4.7)
5: $X \leftarrow X, X_1, ..., X_M$ `with` $X_i \sim N(\texttt{GradientStep}(x, \delta), \varepsilon \times I_d)$
6: $Y \leftarrow f(X)$
7: $x \leftarrow \text{argmin}_{X_i} \{f(X_i)\}$
8: **end while**
9: **return** x

Blackbox Optimization

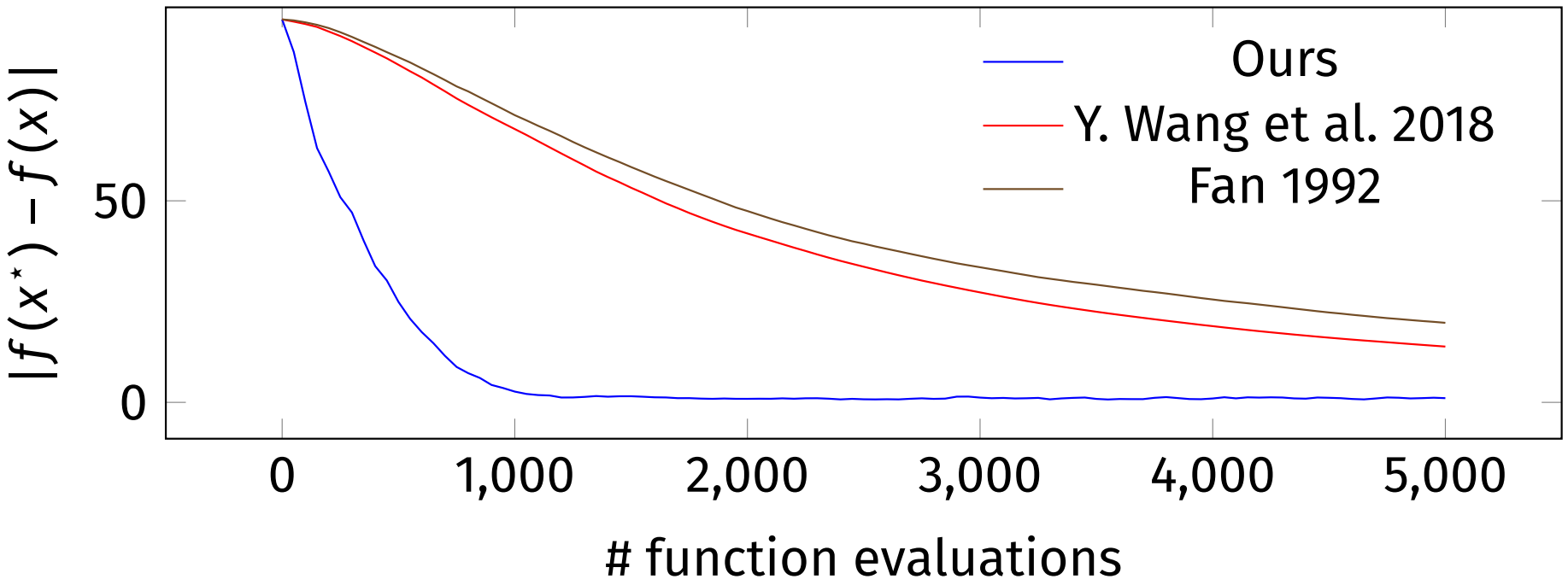
$$x_{n+1} = x_n - \gamma \underbrace{\hat{\nabla}_n f(x_n)}_{\text{Estimated}}$$

Require: x_0: initial guess, f: function $\mathbb{R}^d \mapsto \mathbb{R}$, M: budget

1: $X \leftarrow X_1, ..., X_M$ `with` $X_i \sim N(x_0, \varepsilon \times I_d)$
2: $Y \leftarrow f(X) \coloneqq f(X_1), ..., f(X_M)$
3: **while** `not StoppingCondition` **do**
4: $r, \delta \leftarrow$ `estimated gradient at` x `w.r.t` X, Y `using` eq. (4.7)
5: $X \leftarrow X, X_1, ..., X_M$ `with` $X_i \sim N(\texttt{GradientStep}(x, \delta), \varepsilon \times I_d)$
6: $Y \leftarrow f(X)$
7: $x \leftarrow \text{argmin}_{X_i} \{f(X_i)\}$
8: **end while**
9: **return** x

Blackbox Optimization

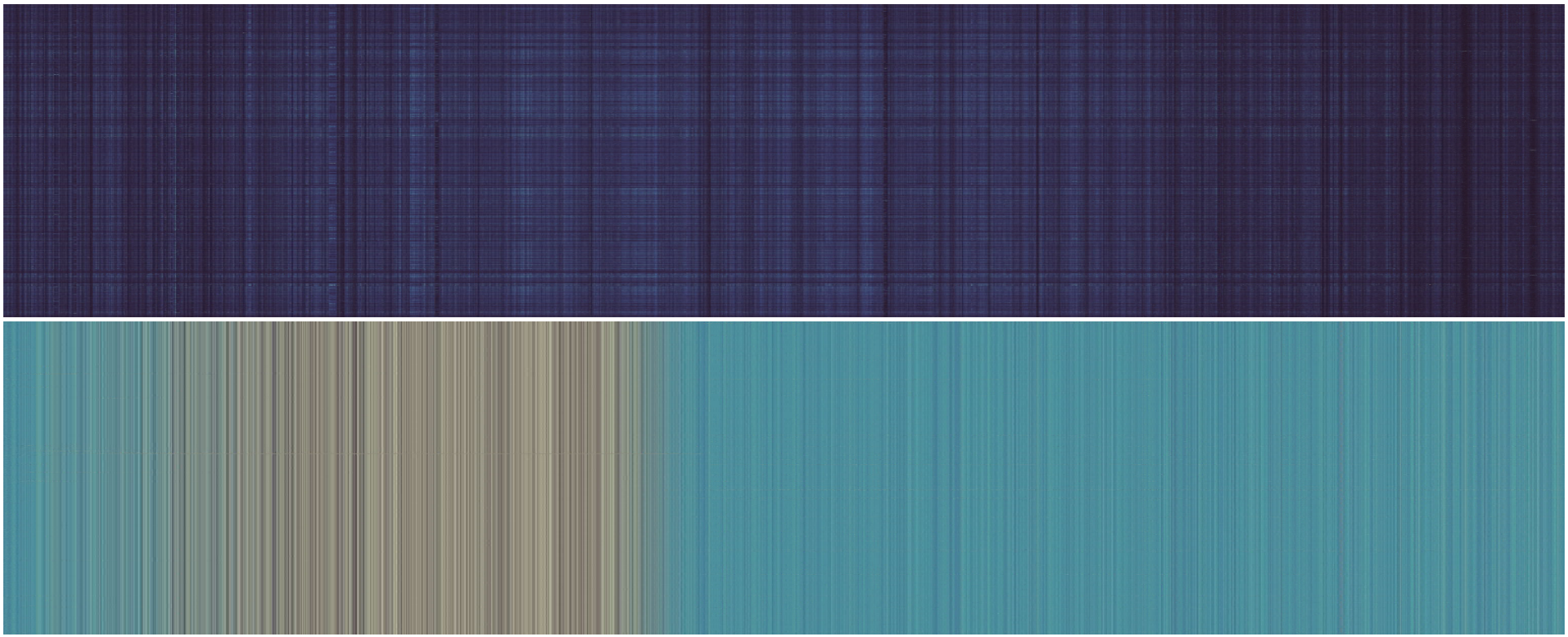
$$f(x) = 100 \sum_{i=1}^{d-1} (x_{i+1} - x_i)^2 + (x_i - 1)^2.$$



Our method reuses past computations to improve performance.

Survival Gradients

$$\underset{(\alpha,\beta)\in\mathbb{R}^{d+1}}{\operatorname{argmin}} \sum_{i\in\hat{i}_k(x)} \frac{\delta_i}{\hat{S}_C(T_i- \mid X_i)} \left(T_i - \alpha - \beta^\intercal (X_i - x)\right)^2 + \lambda \|\beta\|_1,$$



Conclusion & Perspectives

Conclusion & Perspectives

- ERM on censored data after reweighting.
  ⇒ Same rates and non-asymptotic, uniform guarantees.
- Adapt flexible neural normalizing flows to survival.
  ⇒ Can be used with bayesian setting / unstructured data.
- Deal with high-dimension by gradient variable selection.
  ⇒ Local selection of variables.

Conclusion & Perspectives

- ERM on censored data after reweighting.
  ⇒ Same rates and non-asymptotic, uniform guarantees.
- Adapt flexible neural normalizing flows to survival.
  ⇒ Can be used with bayesian setting / unstructured data.
- Deal with high-dimension by gradient variable selection.
  ⇒ Local selection of variables.

But still many perpectives:

- Going from X to $X(t)$.
- Doubly-Robust ERM ?
- Scaling survival flows, deploying on textual data.
- Local Linear rates depend on d, can it depend on manifold dimension $m \ll d$?

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.

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.

— (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.

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.

- Andersen, Per Kragh, Ørnulf Borgan, Richard D. Gill, and Niels Keiding (1993). *Statistical Models Based on Counting Processes*. Springer Series in Statistics. New York, NY: Springer US. DOI: 10.1007/978-1-4612-4348-9 (cit. on pp. 10, 11).
- 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. 42).
- Aswani, Anil, Peter Bickel, and Claire Tomlin (Feb. 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 pp. 46, 81).
- Avesani, Renzo G, Kexue Liu, and Alin Mirestean (2006). "Review and Implementation of Credit Risk Models of the Financial Sector Assessment Program". P. 35 (cit. on p. 6).
- Basel Committee (2019). *Scope and Definitions*. P. 41 (cit. on pp. 4, 5).

- Bielecki, Tomasz R. and Marek Rutkowski (2004). *Credit Risk: Modeling, Valuation and Hedging*. Springer Finance. Berlin, Heidelberg: Springer Berlin Heidelberg. DOI: 10.1007/978-3-662-04821-4 (cit. on p. 6).
- Bücher, Axel, Anouar El Ghouch, and Ingrid Van Keilegom (2014). "Single-Index Quantile Regression Models for Censored Data". In: *Advances in Contemporary Statistics and Econometrics*. Ed. by Abdelaati Daouia and Anne Ruiz-Gazen. Cham: Springer International Publishing, pp. 177–196. DOI: 10.1007/978-3-030-73249-3_10 (cit. on p. 46).
- Cao, Yang, Shengtai Li, Linda Petzold, and Radu Serban (Jan. 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 pp. 39, 40, 80).

- 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 p. 38).
- 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. 19).
- 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 p. 12).
- Dabrowska, Dorota Maria (1986). "Nonparametric Regression with Censored Survival Time Data". 81 (cit. on p. 19).
- — (1988). "Kaplan-Meier Estimate on the Plane". *The Annals of Statistics* 16.4, pp. 1475–1489 (cit. on p. 19).

- Dabrowska, Dorota Maria (Sept. 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 p. 19).
- Dalalyan, Arnak S., Anatoli Juditsky, and Vladimir Spokoiny (Aug. 2008). “A New Algorithm for Estimating the Effective Dimension-Reduction Subspace”. *Journal of Machine Learning Research* 9, pp. 1647–1678 (cit. on pp. 48, 49).
- Delyon, Bernard and François Portier (Nov. 2016). “Integral Approximation by Kernel Smoothing”. *Bernoulli* 22.4, pp. 2177–2208. DOI: 10.3150/15-BEJ725 (cit. on p. 29).
- Durkan, Conor, Artur Bekasov, Iain Murray, and George Papamakarios (June 10, 2019). *Neural Spline Flows*. arXiv: 1906.04032 [cs, stat]. URL: http://arxiv.org/abs/1906.04032 (cit. on pp. 35, 36).

- 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. Berlin, Heidelberg: Springer, pp. 21–57. DOI: 10.1007/978-3-642-33483-2_2 (cit. on pp. 4, 5, 42).
- Fan, Jianqing (Dec. 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. 47, 55).
- — (Mar. 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 pp. 47–49).
- Fan, Jianqing and Irene Gijbels (Mar. 1, 1996). *Local Polynomial Modelling and Its Applications: Monographs on Statistics and Applied Probability 66*. CRC Press. 362 pp. Google Books: BM1ckQKCXP8C (cit. on pp. 48, 49).

- Fleming, T. and D. Harrington (1991). “Counting Processes and Survival Analysis”. In: DOI: 10.2307/2290673 (cit. on pp. 10, 11).
- Giné, Evarist and Armelle Guillou (July 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 p. 29).
- Giné, Evarist and Hailin Sang (Aug. 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. 29).
- Hristache, Marian, Anatoli Juditsky, Jörg Polzehl, and Vladimir Spokoiny (Dec. 2001). “Structure Adaptive Approach for Dimension Reduction”. *Annals of Statistics* 29.6, pp. 1537–1566. DOI: 10.1214/aos/1015345954 (cit. on p. 46).

- Hristache, Marian, Anatoli Juditsky, and Vladimir Spokoiny (May 1998). *Direct Estimation of the Index Coefficients in a Single-Index Model*. report. INRIA (cit. on p. 46).
- Jiang, Heinrich (July 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. 48, 49, 83).
- Kingma, Diederik P. and Prafulla Dhariwal (July 10, 2018). *Glow: Generative Flow with Invertible 1x1 Convolutions*. arXiv: 1807.03039 [cs, stat]. URL: http://arxiv.org/abs/1807.03039 (cit. on pp. 35, 36).
- Lopez, Jose A and Marc R Saidenberg (1999). "Evaluating Credit Risk Models". P. 23 (cit. on p. 6).

- Major, Péter (Mar. 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. 29).
- Merton, Robert C. (1974). “On the Pricing of Corporate Debt: The Risk Structure of Interest Rates*”. *The Journal of Finance* 29.2, pp. 449–470. DOI: 10.1111/j.1540-6261.1974.tb03058.x (cit. on p. 6).
- Mukherjee, Sayan, Qiang Wu, and Ding-Xuan Zhou (Feb. 2010). “Learning Gradients on Manifolds”. *Bernoulli* 16.1, pp. 181–207. DOI: 10.3150/09-BEJ206. arXiv: 1002.4283 (cit. on pp. 46, 81).
- Papamakarios, George, Theo Pavlakou, and Iain Murray (June 14, 2018). *Masked Autoregressive Flow for Density Estimation*. arXiv: 1705.07057 [cs, stat]. URL: http://arxiv.org/abs/1705.07057 (cit. on pp. 35, 36).

- Rackauckas, Christopher, Yingbo Ma, Vaibhav Dixit, et al. (Dec. 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. 39, 40, 80).
- Rackauckas, Christopher, Yingbo Ma, Julius Martensen, et al. (Aug. 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. 80).
- Rackauckas, Christopher and Qing Nie (2017). "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. 80).

- Rezende, Danilo and Shakir Mohamed (June 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. 35, 36).
- 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. 43).
- 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 p. 19).
- — (1993b). “Consistent Estimation under Random Censorship When Covariables Are Present”. *Journal of Multivariate Analysis* 45.1. DOI: 10.1006/jmva.1993.1028 (cit. on p. 19).
- — (Apr. 1995a). “The Central Limit Theorem Under Random Censorship”. *Annals of Statistics* 23.2, pp. 422–439. DOI: 10.1214/aos/1176324528 (cit. on p. 19).

- Stute, Winfried (1995b). “The Statistical Analysis of Kaplan-Meier Integrals”. *Analysis of censored data* 27, pp. 231–254. DOI: 10.1214/lnms/1215452223 (cit. on p. 19).
- — (1996). “Distributional Convergence under Random Censorship When Covariables Are Present”. *Scandinavian Journal of Statistics* 23.4, pp. 461–471 (cit. on p. 19).
- — (2003). “Kaplan-Meier Integrals”. *Handbook of Statistics* 23.03, pp. 87–104. DOI: 10.1016/S0169-7161(03)23005-4 (cit. on p. 19).
- Stute, Winfried and J.-L. Wang (Sept. 1993). “The Strong Law under Random Censorship”. *The Annals of Statistics* 21.3, pp. 1591–1607. DOI: 10.1214/aos/1176349273 (cit. on p. 19).
- Van Keilegom, Ingrid (1998). “Nonparametric Estimation of the Conditional Distribution in Regression with Censored Data”. (Cit. on pp. 10, 11).

- Wang, Yining, Simon Du, Sivaraman Balakrishnan, and Aarti Singh (Mar. 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 p. 55).
- Wehenkel, Antoine and Gilles Louppe (Feb. 12, 2021). *Graphical Normalizing Flows*. arXiv: 2006.02548 [cs, stat]. URL: http://arxiv.org/abs/2006.02548 (cit. on pp. 35, 36).

Backup

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 $h > 0$,

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

Uniform control of U-statistics

Let $\xi_1, \xi_2, ...$ be i.i.d. r.v.s. H a class of functions on S^k uniformly bounded such that H is of vc type with constants (A, v) and envelope G. Set $\sigma^2(H) = V[H(\xi_1, ..., \xi_k)]$ and $\mathbb{E}[H(\xi_1, ..., \xi_k)|\xi_1, ..., \xi_{j-1}, \xi_{j+1}, ..., \xi_k] = 0$. a.s.

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

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

provided that

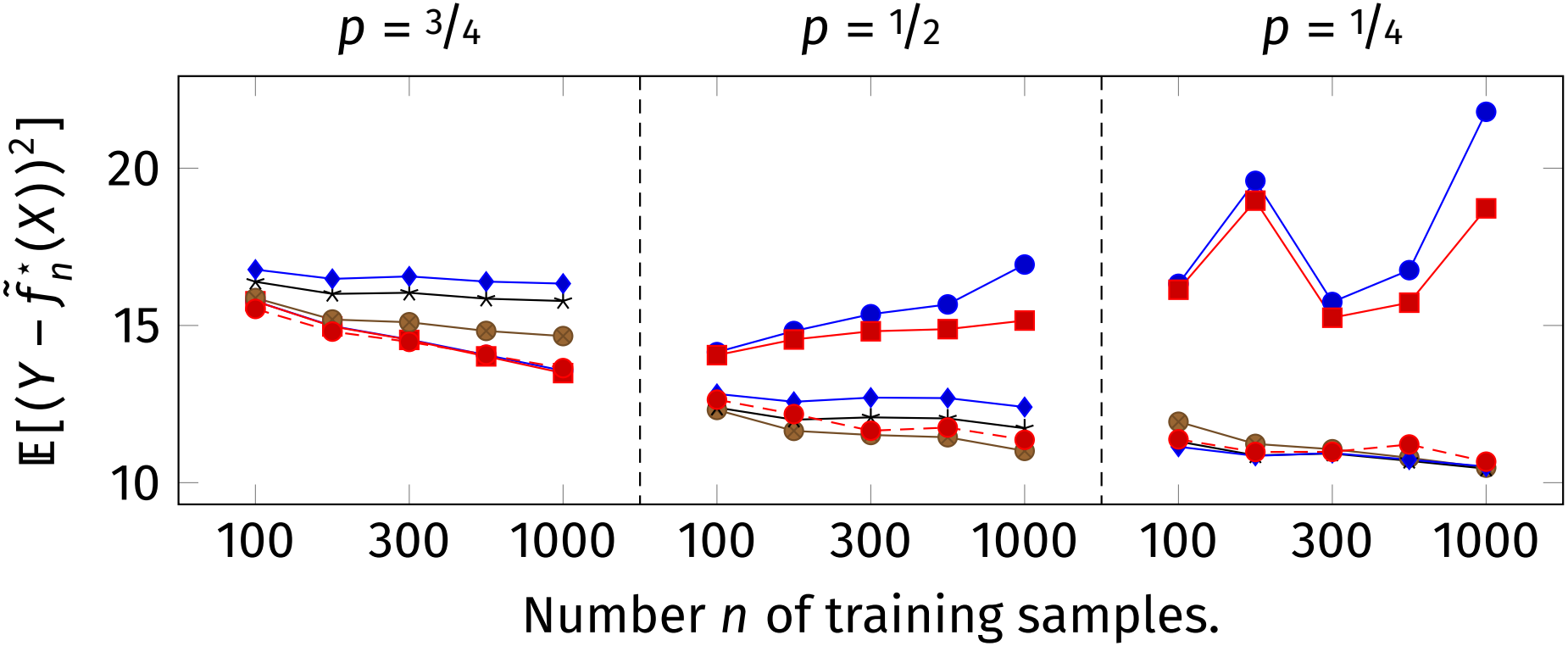
$$\|G\|_\infty^2 \left(C_1^{2/k} \log\left(\frac{2\|G\|_\infty}{\sigma} \right) + \frac{\log(C_2/\varepsilon)}{C_3} \right) \leq n\sigma^2,$$

$$\sup_{H \in H} \sigma^2(H) \leq \sigma^2 \leq \|G\|_\infty^2.$$

Experimental Results

We do not use exactly the estimator studied in order to prevent problems where $\hat{S}_{C,n} = 0$.

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



Experimental Results

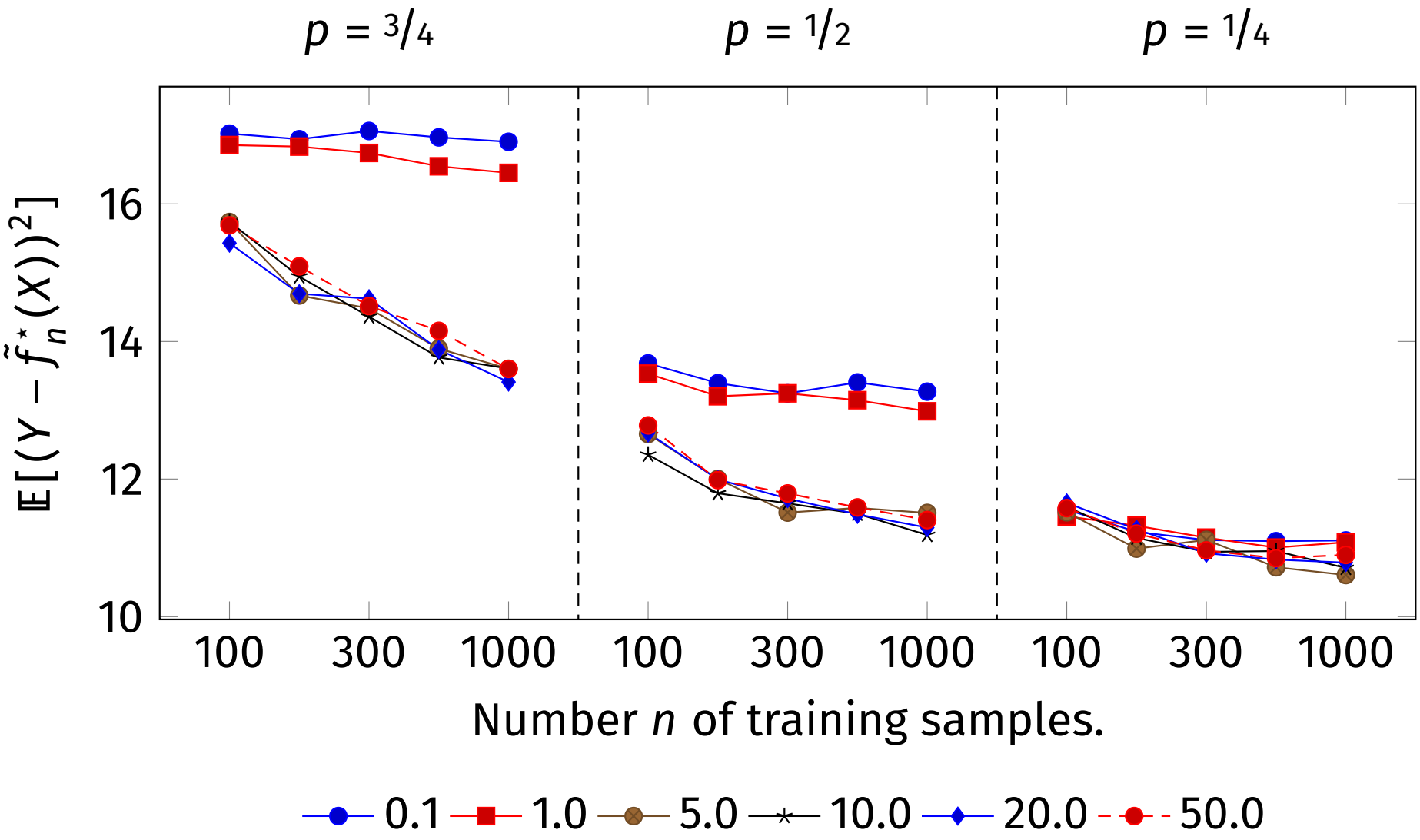


Figure: Prediction error $\mathbb{E}[(Y - \tilde{f}_n(X))^2]$ for varying bandwidth h, with F a

Experimental Results

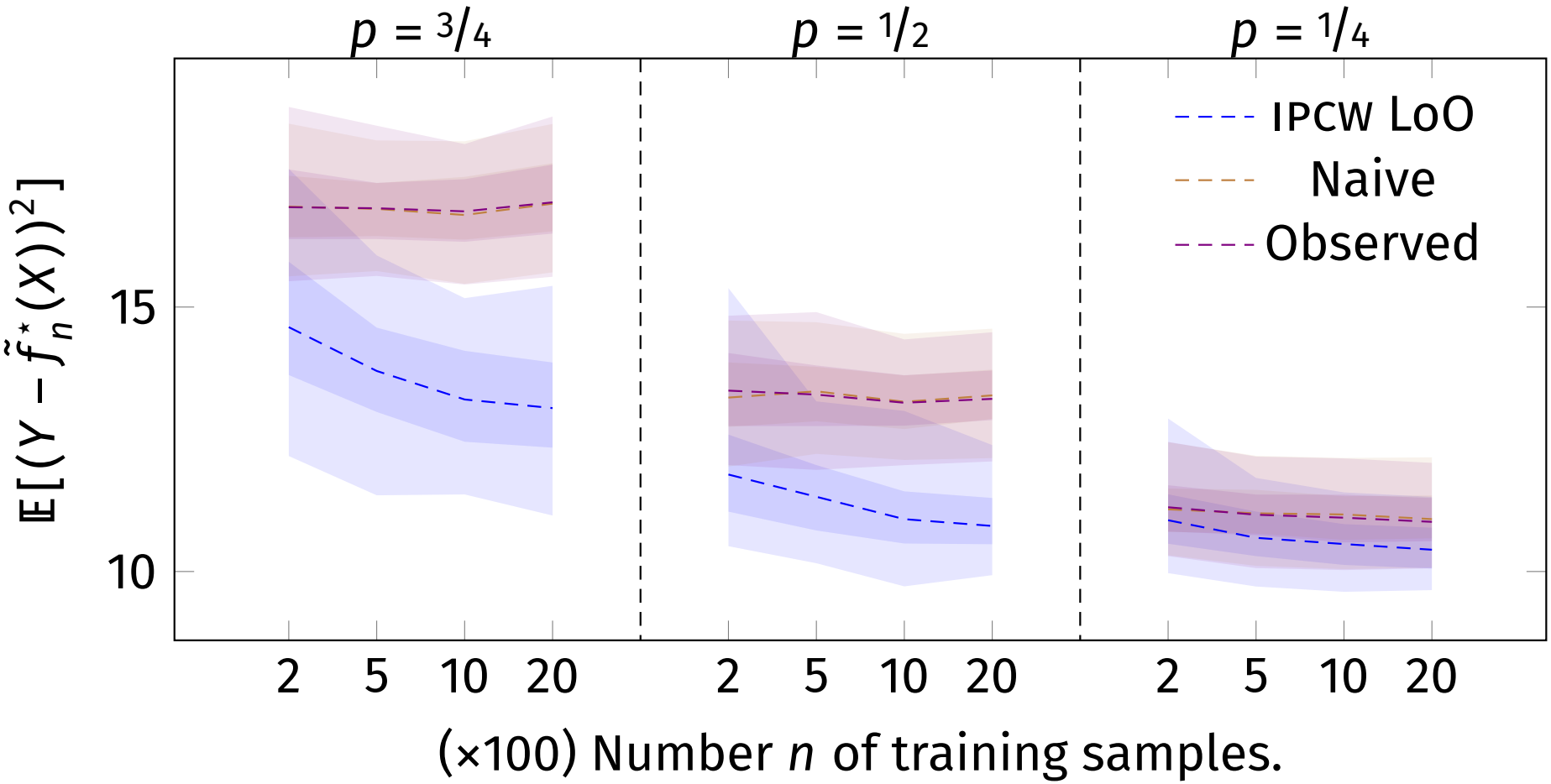


Figure: Prediction error $\mathbb{E}[(Y - \tilde{f}_n(X))^2]$ when F is the class of affine functions, choosing the IPCW LOO risk estimator and the Cox model of eq. (2.28) for generating the data.The curves correspond to different floors b

Experimental Results: TCGA Cancer Dataset

	IPCW		Naive		Observed	
Method	L^2	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

Optimizing with Adjoints

Computing the quantities is not enough. We need to be able to differentiate them. Actually possible "cheaply" by means of the adjoint method (Cao et al. 2003; Rackauckas, Ma, Martensen, et al. 2020; Rackauckas and Nie 2017; Rackauckas, Ma, Dixit, et al. 2018).

$$L(\mathbf{u}, \theta) = \int_{t_0}^{T} l(\mathbf{u}(t, \theta), \theta) \, \mathrm{d}t.$$

We can then form the adjoint state

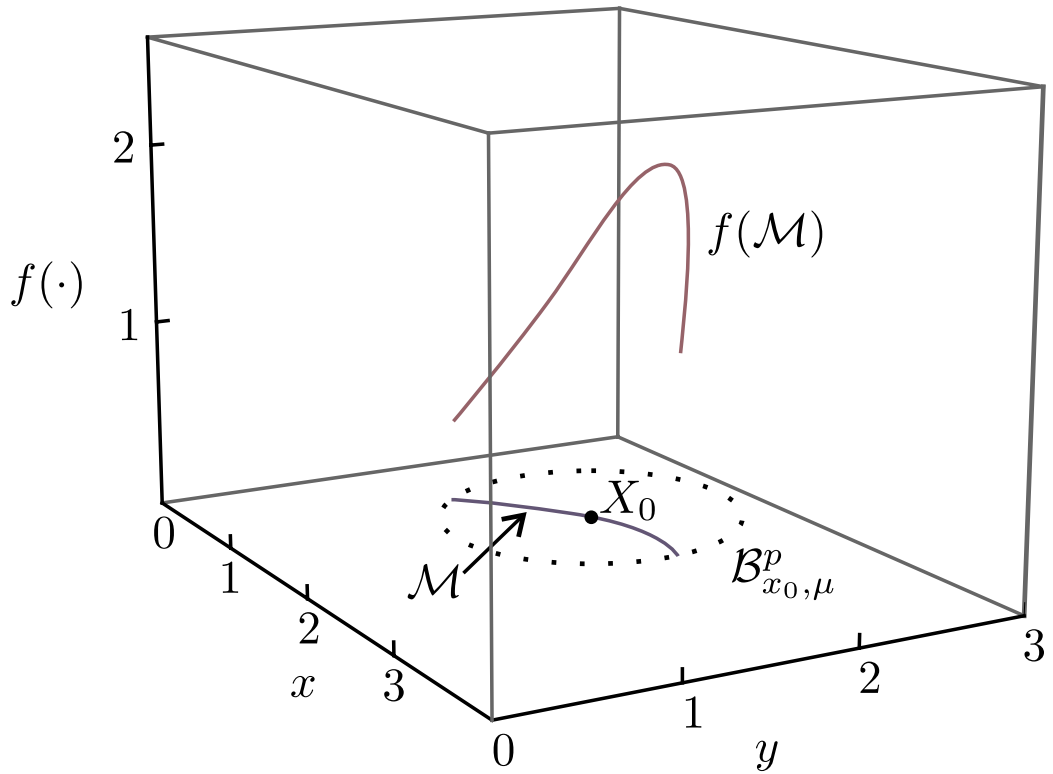
$$\begin{aligned} \frac{\partial \lambda}{\partial t} &= \frac{\partial l}{\partial u}(\mathbf{u}(t, \theta), \theta) - \lambda(t) \frac{\partial p}{\partial u}(t, \mathbf{u}(t, \theta), \theta), \\ \lambda(T) &= 0, \end{aligned}$$

such that

$$\frac{\partial L}{\partial \theta} = \int_{t_0}^{T} \lambda(t) \frac{\partial p}{\partial \theta}(\mathbf{u}(t, \theta), \theta) + \frac{\partial l}{\partial \theta}(\mathbf{u}(t, \theta), \theta) \, \mathrm{d}t + \lambda(t_0) \frac{\partial \mathbf{u}}{\partial \theta}(t_0, \theta).$$

Local Linear Estimator: Manifold

Even possible to ensure a manifold hypothesis (Mukherjee, Wu, and Zhou 2010; Aswani, Bickel, and Tomlin 2011).



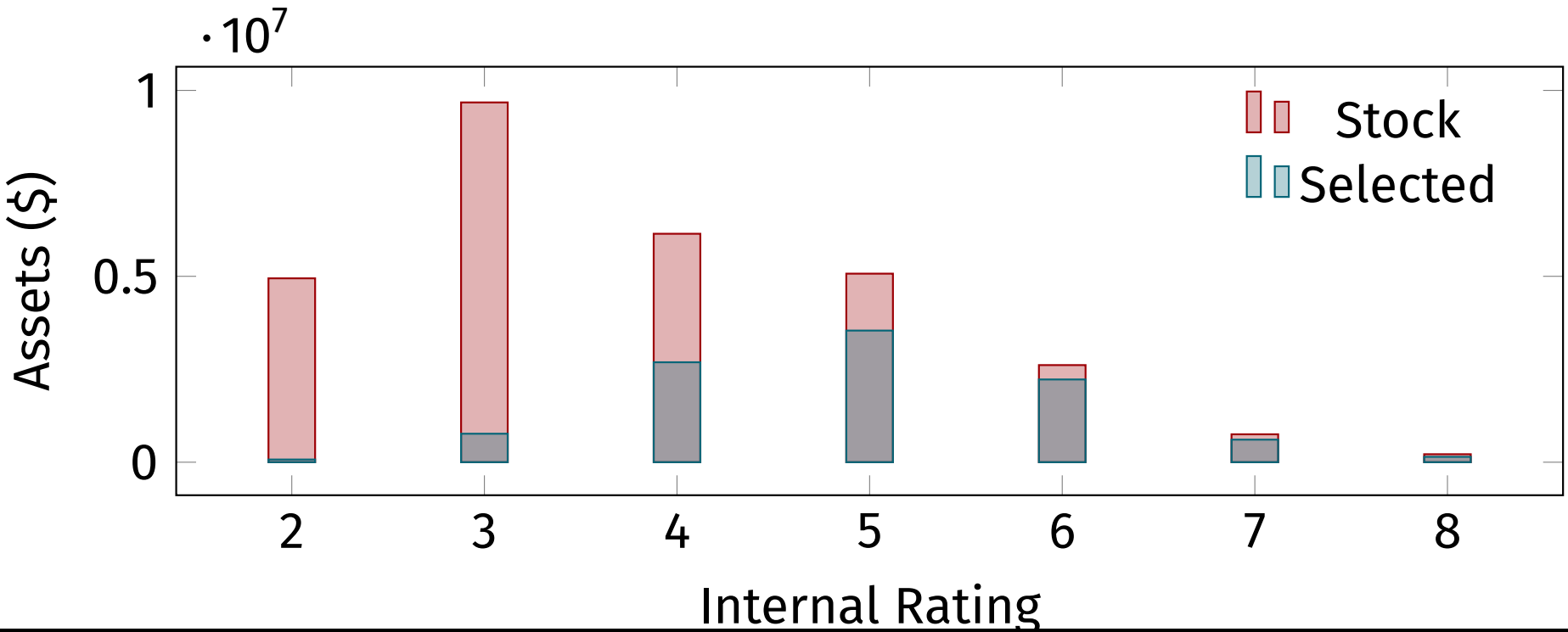
$$\hat{\beta} = \arg\min_{\tilde{\beta}} \left\{ h^p \left\| W_{x_0}^{1/2} \left(Y - X_{x_0} \tilde{\beta} \right) \right\|_2^2 + \lambda_n \left\| \hat{P}_n \cdot \tilde{\beta} \right\|_2^2 + \mu_n \sum_{j=1}^{p} \frac{1}{\hat{w}_j^{\gamma}} \left| \tilde{\beta}_j \right| \right\}$$

Hierarchical Modelling of Defaults

Many models on similar distributions: Hierarchical Pooling.
Bayesian Models / MCMC needs:

- Sampleable
- Tractable Likelihood

$$\theta_i \sim N(\mu, \Sigma)$$
$$Y \mid X, \text{pop } i \sim \text{NF}(\theta_i, X).$$



Bounds on the Regressor

Ausset, Clémençon, and Portier 2021b, Theorem 4.2

Suppose that the previous mild assumptions 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 $1-\delta$:

$$\left|\hat{r}_k(x) - r(x)\right| \leq \sqrt{\frac{2\sigma^2 \log(4/\delta)}{k}} + L_1\left(\frac{2k}{n b_f V_d}\right)^{1/d},$$

where $V_d = \int \mathbb{1}_{B(0,1)}(x)\,\mathrm{d}x$ denotes the volume of the unit ball.

Weaker condition on k than Jiang 2019. Minimax rate.