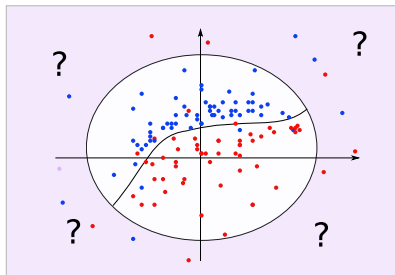


Classification in extreme regions

joint work

Stéphan Cléménçon, Hamid Jalalzai, Anne Sabourin
Télécom ParisTech, Paris, France

February 2019, Ecole Polytechnique, Paris.



Outline

Motivations

Background on classification and extremes, problem statement

Optimal elements for extreme classification

Empirical Risk minimization

Experiments

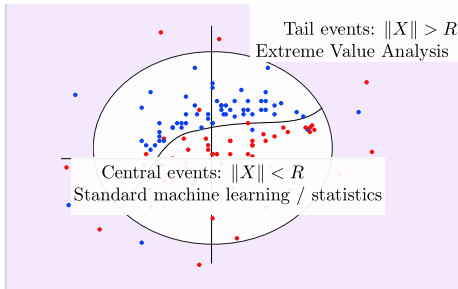
Sketches of proofs

Classification in Extreme regions: why?

- For **risk management, anomaly detection**: monitoring a system based on multivariate data $X \in \mathbb{R}^d$, with unknown state label $Y \in \{-1, +1\}$ (e.g. normal versus abnormal)
- Many instances of interest may be located **in the tails** of the distribution
 - Among network reports with unusually large connection numbers, which ones are **attacks** ?
 - Among unusually large transactions: which ones are **frauds** ?
 - Among tweets with unusually large repetitions of words which are otherwise rare, which ones are related to **buzzes** ?
- → a classification problem with focus on the tail.

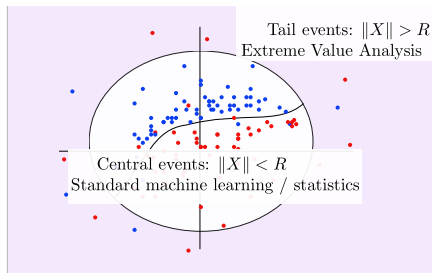
Issue

- **Standard machine learning** tools for classification (SVM, Neural networks, random forests ...) are based on **empirical risk minimisation**
→ well suited for classifying points from the bulk (= central region)
- **Extremes do not contribute much to the empirical error** (proportion $k/n \ll 1$)
→ Extreme points may be hard to classify (no kidding)



A novel methodology for classification of the most extreme points ($\|X\| > t, t \rightarrow \infty$) of a dataset

- Workflow:



1. Pick your favorite classifier (random forest, logistic regression, deep neural network, ...)
2. Train it on a fraction of your data (those with the largest norm)
3. For a new (unlabelled) point x_{new} to be classified:
 - If $\|x_{new}\|$ is small, use an of-the-shelf ML classifier
 - If $\|x_{new}\|$ is large, use the classifier dedicated to extremes.

Outline

Motivations

Background on classification and extremes, problem statement

Optimal elements for extreme classification

Empirical Risk minimization

Experiments

Sketches of proofs

Classification *via* empirical risk minimization (standard approach)

- *i.i.d.* data (X_i, Y_i) , $X_i \in \mathbb{R}^d$, $Y_i \in \{-1, +1\}$, $i = 1, \dots, n$
- **Goal:** train a classifier $g : \mathbb{R}^d \rightarrow \{-1, +1\}$ with low classification risk

$$R(g) = \mathbb{P}(g(X) \neq Y)$$

- Since the law $\mathcal{L}(X, Y)$ is unknown, use the **empirical risk** instead

$$R_n(g) = \frac{1}{n} \sum_{i=1}^n \mathbf{1}\{g(X_i) \neq Y_i\}$$

Classification *via* empirical risk minimization (standard approach)

- **Empirical risk minimizer:** in a class \mathcal{G} 'not too complex', define

$$\hat{g} = \operatorname{argmin}_{g \in \mathcal{G}} R_n(g)$$

- Statistical guarantees e.g. if \mathcal{G} has finite VC dimension $V_{\mathcal{G}}$: with probability $1 - \delta$,

$$R(\hat{g}) - R(g^*) \leq C \sqrt{\frac{V_{\mathcal{G}} \log n + \log(1/\delta)}{n}}$$

where $g^* = \operatorname{argmin}_{g \in \mathcal{G}} R(g)$, $g^*(X) = 2\mathbf{1}\{\eta(X) > 1/2\} - 1$ and $\eta(x) = \mathbb{P}(Y = +1|x)$.

Classification risk on extreme regions

- Since we are interested in classifying extreme points, define

$$L_t(g) = \mathbb{P}(g(X) \neq Y \mid \|X\| > t) \quad (\text{classification risk above level } t)$$

- For extremes (out-of-sample), the quantity to be minimized is

$$L_\infty(g) = \limsup_{t \rightarrow \infty} L_t(g). \quad (\text{asymptotic risk in the extremes})$$

Questions

Under appropriate heavy-tail assumptions,

- Is there a minimizer for L_∞ ?
- Can it be characterized in terms of the limiting distribution of (X, Y) given that $\|X\|$ is large ?
- If \hat{g}_t is an empirical minimizer of L_t , can we bound the excess risk $L_\infty(\hat{g}_t) - L_\infty^*$, as $t, n \rightarrow \infty$?

Outline

Motivations

Background on classification and extremes, problem statement

Optimal elements for extreme classification

Empirical Risk minimization

Experiments

Sketches of proofs

Answer to one easy question

- **Is there a minimizer for L_∞ ?**

Answer to one easy question

- **Is there a minimizer for L_∞ ?**

YES: for any classifier g , $L_t(g) \geq L_t(g^*)$ where g^* is the Bayes classifier. Taking the lim sup as $t \rightarrow \infty$:

$$L_\infty(g^*) \leq L_\infty(g)$$

Thus

$$g^* \in \operatorname{argmin}_g L_\infty(g).$$

(and $L_\infty^* = L_\infty(g^*)$).

Answer to one easy question

- **Is there a minimizer for L_∞ ?**

YES: for any classifier g , $L_t(g) \geq L_t(g^*)$ where g^* is the Bayes classifier. Taking the lim sup as $t \rightarrow \infty$:

$$L_\infty(g^*) \leq L_\infty(g)$$

Thus

$$g^* \in \operatorname{argmin}_g L_\infty(g).$$

(and $L_\infty^* = L_\infty(g^*)$).

- **Problem solved?**

NO: Does not provide guarantees for $L_\infty(\hat{g})$.

→ We need additional assumptions.

Multivariate heavy tail data

- **Regular variation** (see Resnick 87, 2007, Hult & Lindskog 2006) :

$X \in \mathbb{R}_+^d$, random vector. $\exists \alpha > 0$ (regular variation index) and μ (exponent measure) on $\mathbb{R}_+^d \setminus \{0\}$ s.t.

$$t^\alpha \mathbb{P}(X \in tA) \xrightarrow[t \rightarrow \infty]{} \mu(A)$$

for all $A \subset \mathbb{R}_+^d$ with $\mu(\partial A) = 0$.

- Consequence: Exponent measure μ is homogeneous,

$$\mu(tA) = t^{-\alpha} \mu(A).$$

- **Intuition:** μ determines the distribution of extremes.
For $B = tA$, far from 0,

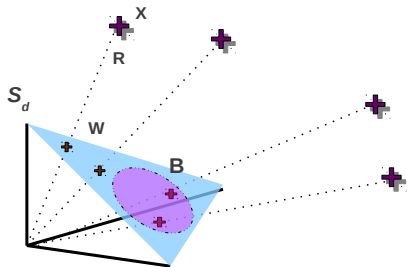
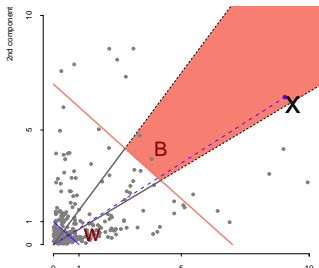
$$\mathbb{P}(X \in B) \simeq \mu(B)$$

Angular measure

(de Haan, Resnick, 77)

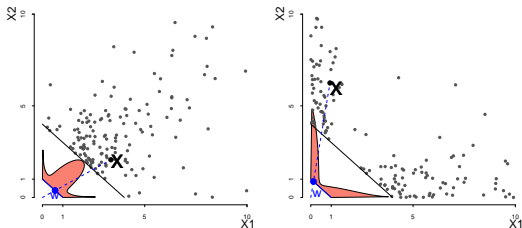
- Polar coordinates: $R = \|X\|_1 = \sum_{j=1}^d X_j$, $W = \frac{X}{R}$.
- $W \in \text{simplex } S_d = \{w : w_j \geq 0, \sum_j w_j = 1\}$.
- Angular measure: $\Phi(B) := \mu\{tw, w \in B, t \geq 1\}$ (cone with base B)
- Above large thresholds r_0 :

$$\mathbb{P}(R > r, W \in B \mid R \geq r_0) \simeq \left(\frac{r_0}{r}\right)^\alpha \Phi(B)/\Phi(S_d)$$



Angular distribution

- $\Phi (+ \alpha)$ rules the joint distribution of extremes



- **Intuition:** The optimal classifier in extreme regions should be related with Φ^+ , Φ^- (angular measures for the positive and negative instances respectively)

Assumption 1: standard heavy tailed class distribution

- $\mathcal{L}(X|Y = +1)$ and $\mathcal{L}(X|Y = -1)$: regularly varying with
 - same index $\alpha = 1$,
 - exponent measures μ^+, μ^- ,
 - angular measures Φ^+, Φ^- , i.e.

$$t\mathbb{P}(t^{-1}X \in A | Y = \sigma 1) \xrightarrow[t \rightarrow \infty]{} \mu_\sigma(A), \quad \sigma \in \{-, +\},$$

- Consequence: X is RV with exponent measure

$$\mu = p\mu^+ + (1 - p)\mu^-,$$

where $p = \mathbb{P}(Y = +1)$

Remarks on Assumptions 1

- **Remark 1:** if $\alpha^+ < \alpha^-$, the positive class has heavier tails, and the trivial classifier $g(x) = 1$ is asymptotically optimal.

Remarks on Assumptions 1

- **Remark 1:** if $\alpha^+ < \alpha^-$, the positive class has heavier tails, and the trivial classifier $g(x) = 1$ is asymptotically optimal.
- **Remark 2:** if $\alpha \neq 1$, one can standardize X using an empirical transformation based on ranks to get back to $\alpha = 1$.

Asymptotic distribution of (X, Y)

- Consider a 'pseudo' (not observed) random pair

$$(X_\infty, Y_\infty) \in (\mathbb{R}_+^d \cap \{\|x\| \geq 1\}) \times \{-1, +1\}$$

distributed as the limit law of (X, Y) given that $\|X\| > t$, i.e.

$$\begin{aligned} \mathbb{P}(Y_\infty = +1) &= \lim_t \mathbb{P}(Y = 1 \mid \|X\| > t) \\ &:= p_\infty \end{aligned}$$

$$\mathbb{P}(X_\infty \in A, Y_\infty = +1) = \lim_t \mathbb{P}(X \in tA, Y = +1 \mid \|X\| > t)$$

Bayes classifier relative to extremes

- One can show that the regression function for (X_∞, Y_∞) writes

$$\eta_\infty(x) = \mathbb{P}(Y_\infty = +1 \mid X = x) = \frac{p\varphi^+(w)}{p\varphi^+(w) + (1-p)\varphi^-(w)}$$

The regression function for the pair (X_∞, Y_∞) depends on $w = x/\|x\|$ only!

Bayes classifier relative to extremes

- One can show that the regression function for (X_∞, Y_∞) writes

$$\eta_\infty(x) = \mathbb{P}(Y_\infty = +1 \mid X = x) = \frac{p\varphi^+(w)}{p\varphi^+(w) + (1-p)\varphi^-(w)}$$

The regression function for the pair (X_∞, Y_∞) depends on $w = x/\|x\|$ only!

- Reminder: the optimal classifier for the classical risk $\mathbb{P}(g(X_\infty) \neq Y_\infty)$ is the Bayes classifier $g_\infty(x) = 2\mathbf{1}\{\eta_\infty(x) > 1/2\} - 1$,

The Bayes classifier g_∞ for the pair (X_∞, Y_∞) depends on $w = x/\|x\|$ only!

The Bayes classifier for extremes is asymptotically optimal

- **Assumption 2:** $\sup_{\{x \in \mathbb{R}_+^d : \|x\| \geq t\}} |\eta(x) - \eta_\infty(x)| \xrightarrow[t \rightarrow \infty]{} 0.$
(satisfied in the framework of de Haan, 87, Cai et al. 2011)

The Bayes classifier for extremes is asymptotically optimal

- **Assumption 2:** $\sup_{\{x \in \mathbb{R}_+^d : \|x\| \geq t\}} |\eta(x) - \eta_\infty(x)| \xrightarrow[t \rightarrow \infty]{} 0.$
(satisfied in the framework of de Haan, 87, Cai et al. 2011)

Theorem 1

Under assumptions 1 and 2

- The extreme Bayes classifier and the classical one are asymptotically equivalent in terms of extreme risk: $L_t(g_\infty) - L_t(g^*) \rightarrow 0$
- The minimum asymptotic risk in the extremes is the bayes risk of the limiting distribution: $L_\infty^* = \lim_t L_t(g^*) = L_{(X_\infty, Y_\infty)}(g_\infty)$

- **Consequence:**

There is an optimal classifier for the asymptotic risk L_∞ depending on $w = x/\|x\|$ only.

Outline

Motivations

Background on classification and extremes, problem statement

Optimal elements for extreme classification

Empirical Risk minimization

Experiments

Sketches of proofs

Empirical Risk Minimization for extremes

- Theorem 2 suggest working with classifiers of the kind $g(x) = g(x/\|x\|)$. Let \mathcal{G}_S such a family with finite VC-dimension $V_{\mathcal{G}_S}$.
- For $0 < \tau \ll 1$, let $k = \lfloor n\tau \rfloor \ll n$.
- Empirical risk above level t_τ :

$$\hat{L}_k(g) = \frac{1}{k} \sum_{i=1}^k \mathbf{1}\{g(X_{(i)}) \neq Y_{(i)}\}$$

where $\|X_{(1)}\| \geq \|X_{(2)}\| \geq \dots \|X_{(n)}\|$.

- We consider the classifier $\hat{g}_k = \operatorname{argmin}_{g \in \mathcal{G}_S} \hat{L}_k(g)$
(= output of the training step for classification in extreme regions)

A bound on the excess risk

Set $t_\tau = (1 - \tau)$ quantile of $\|X\|$.

Theorem 2

For $\delta \in (0, 1)$, $\forall n \geq 1$, we have with probability larger than $1 - \delta$:

$$L_{t_\tau}(\hat{g}_k) - L_{t_\tau}^* \leq C \sqrt{\frac{V_{\mathcal{G}_S} \log(1/\delta)}{k}} + O_{\delta, V_{\mathcal{G}_S}}(1/k) + \underbrace{\left\{ \inf_{g \in \mathcal{G}_S} L_{t_\tau}(g) - L_{t_\tau}^* \right\}}_{\text{BIAS}(\tau)},$$

where C is a universal constant independent from n , τ and δ .

Corollary

If $\text{BIAS}(\tau) \rightarrow 0$ as $\tau \rightarrow 0$ (possible if \mathcal{G}_S approaches g_∞), then

$$L_\infty(\hat{g}_k) - L_\infty^* \xrightarrow[n \rightarrow \infty]{} 0$$

whenever $k \rightarrow \infty$, $k/n \rightarrow 0$.

Algorithm: ERM for classification of extremes

Input: Training dataset $\mathcal{D}_n = \{(X_1, Y_1), \dots, (X_n, Y_n)\}$, collection \mathcal{G}_S of classifiers on the sphere, $k \leq n$.

Algorithm: ERM for classification of extremes

Input: Training dataset $\mathcal{D}_n = \{(X_1, Y_1), \dots, (X_n, Y_n)\}$, collection \mathcal{G}_S of classifiers on the sphere, $k \leq n$.

- 1 **Standardize.** $\forall i \in \{1, \dots, n\}$, $\hat{V}_i = \hat{T}(X_i)$, where
$$\hat{T}(x) = \left(1 / \left(1 - \hat{F}_j(x_j) \right) \right)_{j=1, \dots, d},$$

Algorithm: ERM for classification of extremes

Input: Training dataset $\mathcal{D}_n = \{(X_1, Y_1), \dots, (X_n, Y_n)\}$, collection \mathcal{G}_S of classifiers on the sphere, $k \leq n$.

- 1 Standardize.** $\forall i \in \{1, \dots, n\}$, $\hat{V}_i = \hat{T}(X_i)$, where
$$\hat{T}(x) = \left(1 / \left(1 - \hat{F}_j(x_j) \right) \right)_{j=1, \dots, d},$$
- 2 Truncate.** Sort the training input observations $\|\hat{V}_{(1)}\| \geq \dots \geq \|\hat{V}_{(n)}\|$, and consider the set of extreme training points $\left\{ (\hat{V}_{(1)}, Y_{(1)}), \dots, (\hat{V}_{(k)}, Y_{(k)}) \right\}$.

Algorithm: ERM for classification of extremes

Input: Training dataset $\mathcal{D}_n = \{(X_1, Y_1), \dots, (X_n, Y_n)\}$, collection \mathcal{G}_S of classifiers on the sphere, $k \leq n$.

- 1 Standardize.** $\forall i \in \{1, \dots, n\}$, $\hat{V}_i = \hat{T}(X_i)$, where $\hat{T}(x) = \left(1 / \left(1 - \hat{F}_j(x_j)\right)\right)_{j=1, \dots, d}$,
- 2 Truncate.** Sort the training input observations $\|\hat{V}_{(1)}\| \geq \dots \geq \|\hat{V}_{(n)}\|$, and consider the set of extreme training points $\left\{(\hat{V}_{(1)}, Y_{(1)}), \dots, (\hat{V}_{(k)}, Y_{(k)})\right\}$.
- 3 Optimize.** Compute a solution $\hat{g}_k(\theta)$ for the problem

$$\min_{g \in \mathcal{G}_S} \frac{1}{k} \sum_{i=1}^k \mathbf{1} \left\{ Y_{(i)} \neq g \left(\hat{W}_{(i)} \right) \right\}$$

Algorithm: ERM for classification of extremes

Input: Training dataset $\mathcal{D}_n = \{(X_1, Y_1), \dots, (X_n, Y_n)\}$, collection \mathcal{G}_S of classifiers on the sphere, $k \leq n$.

- 1 Standardize.** $\forall i \in \{1, \dots, n\}$, $\hat{V}_i = \hat{T}(X_i)$, where $\hat{T}(x) = \left(1 / \left(1 - \hat{F}_j(x_j)\right)\right)_{j=1, \dots, d}$,
- 2 Truncate.** Sort the training input observations $\|\hat{V}_{(1)}\| \geq \dots \geq \|\hat{V}_{(n)}\|$, and consider the set of extreme training points $\left\{(\hat{V}_{(1)}, Y_{(1)}), \dots, (\hat{V}_{(k)}, Y_{(k)})\right\}$.
- 3 Optimize.** Compute a solution $\hat{g}_k(\theta)$ for the problem

$$\min_{g \in \mathcal{G}_S} \frac{1}{k} \sum_{i=1}^k \mathbf{1} \left\{ Y_{(i)} \neq g \left(\hat{W}_{(i)} \right) \right\}$$

Output: The classifier $\hat{g}_k \left(\hat{T}(x) / \|\hat{T}(x)\| \right)$, applicable on the region $\{x : \|\hat{T}(x)\| > \|\hat{V}_{(k)}\|\}$.

Outline

Motivations

Background on classification and extremes, problem statement

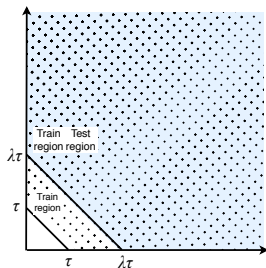
Optimal elements for extreme classification

Empirical Risk minimization

Experiments

Sketches of proofs

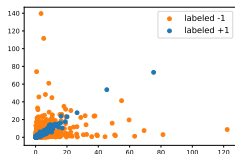
Experimental framework



- Class \mathcal{G}_S : standard ML classifiers (K-NN, random forests).
- Training set for our algorithm: $\{X_i : \|\hat{T}(X_i)\| > \tau\} \rightarrow \hat{g}_k$
- L_∞ : approximated using empirical risk above level $\lambda\tau$ for increasing $\lambda > 1$.
- Comparison: $L_\infty(\hat{g}_k)$ (from our algorithm) versus $L_\infty(\hat{g})$ (classical one, using the whole dataset, no standardization nor truncation)

Simulated data

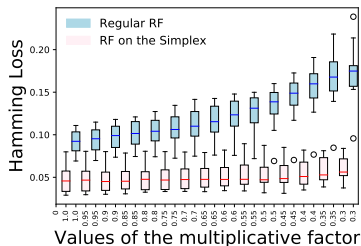
- Both classes from extreme values 'logistic' distributions, dependence parameter $a^+ = 0.1$, $a^- = 0.5$, in dimension $d = 4$.
- Train set: $n_{train} = 10^4$, $k = \sqrt{n} = 100$



Example of 2-classes Logistic dataset, $d = 2$, $n = 10^3$.

Simulated data: results

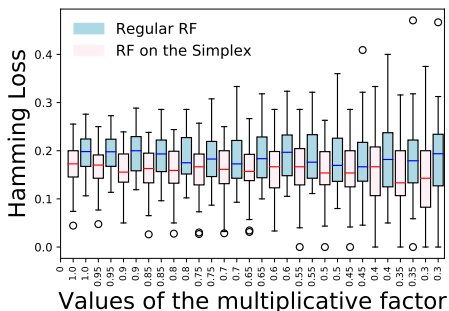
- $n_{test} = 10^3$ test data such that $\|\hat{T}(X_i^{test})\| > t = \|\hat{T}(X_{(k)}^{train})\|$.
- Empirical error on extreme region: using κn_{test} largest test data, $\kappa \in (0, 0.3]$.



Test loss of RF on the simplex and regular RF depending on κ ,
boxplots over 10 experiments

Real world data

- Ecoli dataset (UCI ML database)
- $n = 336, d = 7$.
- Label Y : Protein localization site (8 classes, class labeled 'im' set to 1, other classes set to -1)
- We choose (quite arbitrarily) $k = 100$.



Conclusion

- We provide a general framework for classification of extremes: Our algorithm is rather a meta-algorithm: can be used with any kind of basic classifier as input. Performance in extreme regions is improved compared with standard approaches.
- Perspectives:
 - No dimension reduction (apart from $d \rightarrow d - 1 = \dim(S_d)$): For high dimensional data (e.g. text data) one should apply a dimension reduction device suitable for extreme regions (how ?)
 - Model Selection: may be achieved adding a complexity penalty to the empirical risk
 - Ongoing work: quantify the influence of the empirical standardization (using \hat{F}_j instead of F_j)

some bibliography

- M. Achab, S Cl emen on, A.S., "Max K-armed bandit: On the ExtremeHunter algorithm and beyond." *ECML, 2017*.
- J-J. Cai, J. Einmahl, and L. De Haan. "Estimation of extreme risk regions under multivariate regular variation." *AoS, 2011*
- A. Carpentier and M. Valko. "Extreme bandits". *NIPS, 2014*.
- L. Devroye, L. Gy orfi, and G. Lugosi. "A Probabilistic Theory of Pattern Recognition", *1996*
- N. Goix, A. S., S. Cl emen on. "Learning the dependence structure of rare events: a non-asymptotic study", *.COLT, 2015*
- N. Goix, A. S., and S. Cl emen on. Sparse representation of multivariate extremes with applications to anomaly detection. *JMVA, 2017*
- L. De Haan and S. Resnick. "On regular variation of probability densities". *Stochastic processes and their applications, 1987*.
- **H. Jalalzai, S. Cl emen on, A. S. "On Binary Classification in Extreme Regions." NIPS, 2018.**
- M. Ohannessian and M. Dahleh. "Rare probability estimation under regularly varying heavy tails". *COLT, 2012*

Outline

Motivations

Background on classification and extremes, problem statement

Optimal elements for extreme classification

Empirical Risk minimization

Experiments

Sketches of proofs

Theorem 1: sketch of proof

- $L_t(\mathbf{g}_\infty) - L_t(\mathbf{g}^*) \rightarrow 0$?

Write

$$L_t(\mathbf{g}_\infty) - L_t(\mathbf{g}^*) = \mathbb{E}(f(\eta(X), \mathbf{1}\{\eta(X) < 1/2\} - \mathbf{1}\{\eta_\infty(X) < 1/2\}) \mid \|X\| > t)$$

and use $\eta(x) - \eta_\infty(x) \rightarrow 0$ as $\|x\| \rightarrow \infty$ (Assumption 2)

- $L_t(\mathbf{g}^*) - L_{(X_\infty, Y_\infty)}(\mathbf{g}_\infty) \rightarrow 0$?

Use that $L_t(\mathbf{g}^*) = \mathbb{E}(\min(\eta(X), 1 - \eta(X)) \mid \|X\| > t)$ and combine assumption 2 together with Regular Variation.

Theorem 2: sketch of proof

- Classical Bias/Variance decomposition

$$L_{t_\tau}(\hat{g}_k) - L_{t_\tau}^* \leq 2 \sup_{g \in \mathcal{G}_S} |\hat{L}_k(g) - L_{t_\tau}(g)| + \inf_{g \in \mathcal{G}_S} L_{t_\tau}(g) - L_{t_\tau}^*.$$

- Uniform bound on $|\hat{L}_k(g) - L_{t_\tau}(g)|$?

Use a result from (Goix et al. 15) (concentration inequality on low probability regions):

$$\sup_{A \in \mathcal{A}} \left| \mathbb{P}(Z \in A) - \frac{1}{n} \sum_{i=1}^n \mathbf{1}\{Z_i \in A\} \right| \leq C \left(\sqrt{p} \sqrt{\frac{V_A}{n} \log \frac{1}{\delta}} + \frac{1}{n} \log \frac{1}{\delta} \right)$$

where $p = \mathbb{P}(Z \in \bigcup_{A \in \mathcal{A}} A)$ and C is an absolute constant.