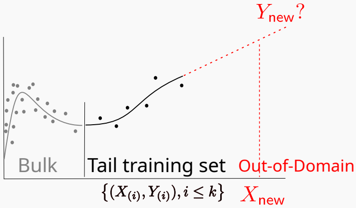
XLASSO: High-Dimensional Regression with Heavy-Tailed Predictors (and targets)

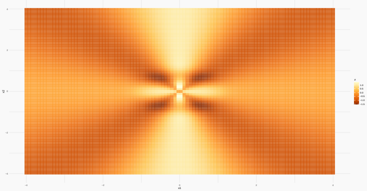
Anne Sabourin

Laboratoire MAP5, Université Paris Cité, CNRS, Paris, France

arXiv:2504.06984

Eindhoven, Eurandom YES workshop, 2025





Outline

Material for this talk

- Final two sections of the survey paper:
  S. Clémençon, A. S., *Weak signals and heavy tails: Machine-learning meets extreme value theory*,
  arXiv preprint arXiv:2504.06984

- Focus of this talk: regression on extreme covariates (Huet et al., 2023)

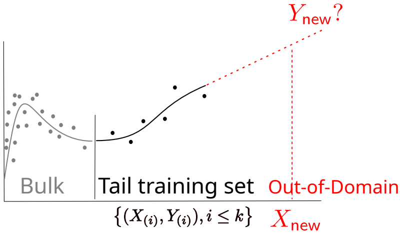
- **What's new today:**
  - High dimension, how to add a penalty term (LASSO) with (some) guarantees
  - Further discussion of (apparently) limitative boundedness assumption on the target

Learning on extreme covariates

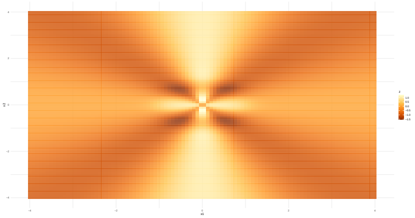
- X: **Heavy tailed random covariates,** Y: **target** to be predicted, $Y \in I = \{-1, 1\}$ (Jalalzai et al., 2018, Binary classification) or $I \subset \mathbb{R}$ (Huet et al., 2023, Regression)

- **Goal:** make acurate prediction in **'crisis scenarios'** where new observed covariable are (unusally) large

- General motivations **Covariate shifts** with climate change, risk management in **worst case events**, ...



1D covariate



2D covariate

Related work

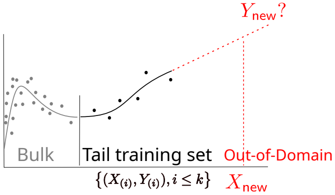
- Already done: stylized setting, no penalty, VC class of sets/functions
  - Jalalzai et al. (2018); Clémençon et al. (2023) for classification, with CV evaluation / hyperparameter selection in Aghbalou et al. (2024).
  - Least squares regression (Huet et al., 2023), continuous target
- Other related:
  - Buriticá and Engelke (2024): similar goal, different choices: Quantile regression, 1D, wide range of tail models.
  - "Cascading extremes" de Carvalho et al. (2025), regression models de Carvalho et al. (2022)
  - Vast literature around the Heffernan-Tawn-Resnick model
- Outside EVT, in ML: "Out-of-Domain generalization", "Transfer learning", " few shots learning", . . .

Outline

Outline

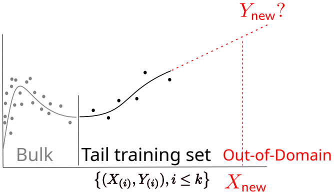
heavy-tailed covariates, but targets?

- First picture in mind



heavy-tailed covariates, but targets?

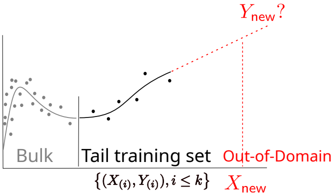
- First picture in mind



- But: boundedness of Y is mathematically (very) convenient for statistical guarantees beyond consistency

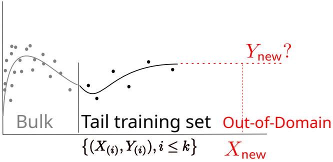
heavy-tailed covariates, but targets?

- First picture in mind



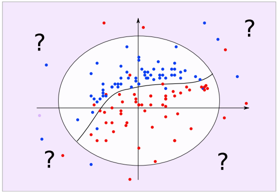
- But: boundedness of Y is mathematically (very) convenient for statistical guarantees beyond consistency

- Let's work with this 'bounded target' assumption (for the moment),



NB: with multivariate X, more complex than this picture.

EVA use cases for binary targets



- Predicting a relative excess (Aghbalou et al., 2024):

$$\tilde{X} = (X_1, \dots, X_{d+1}) \in \mathbb{R}^{d+1} \qquad \text{heavy-tailed,}$$

$$Y = \mathbb{1}\Big\{X_{d+1} \geq c\|(X_1, \dots, X_d)\|\Big\} \quad ; \quad X = (X_1, \dots, X_d)$$

- **focus: (unbounded)** X_{d+1} but binary classification predicts whether **"some component will be large, given that the others are".**

EVA use cases for bounded, real valued targets

- Continuous target:
  $X =$ (temperature, air quality), $Y =$ daily **proportion** of admissions to the pneumology department in a hospital.

- Predicting a **relative value**

$$\tilde{X} = (X_1, \dots, X_{d+1}) \in \mathbb{R}^{d+1} \quad \text{heavy-tailed,}$$

$$Y = X_{d+1}/\|(X_{1:d+1})\| \;\; ; \qquad X = X_{1:d} \qquad \text{(Huet et al., 2023)}$$

or

$$Y = X_{d+1}/\|(X_{1:d})\| \;\; ; \qquad X = X_{1:d} \qquad \text{(Aghbalou et al., 2024)}$$

Again, interest lies in X_{d+1}, but bounded Y carries information
More formalism later

Outline

Conditional risk minimization, asymptotic risk, learning problem (Jalalzai et al., 2018; Huet et al., 2023)

- Conditional risk (finite threshold)

$$R_t(f) = \mathbb{E}\left(c(f(X), Y) \mid \|X\| > t\right).$$

- Out-of-domain risk

$$R_\infty(f) = \limsup_{t\to\infty} R_t(f).$$

- Learning task for out-of-domain extrapolation:

Minimize $R_\infty(f)$ over $f \in \mathcal{F}$ a class of prediction functions, based on i.i.d. data $(X_i, Y_i)_{i\le n} \sim (X, Y)$

Stability assumptions for extrapolation

- Bounded target $|Y| < M$
- Main assumption: "One component regular variation" or "regular variation with respect to the covariate":

$$\left[(t^{-1}X, Y) \quad \middle| \quad \|X\| > t\right] \xrightarrow[t\to\infty]{w} (X_\infty, Y_\infty)$$

- Additional stability assumption

$$\mathbb{E}\left(|f^*(X) - f^*_\infty(X)| \;\middle|\; \|X\| > t\right) \to 0 \tag{1}$$

- Satisfied under (classical) regular variation of densities, similar to De Haan and Resnick (1987); Cai et al. (2011)

Angular regression function in the tails

- Let $\Theta_\infty = \|X_\infty\|^{-1} X_\infty$. Then by homogeneity

$$(Y_\infty, \Theta_\infty) \perp\!\!\!\perp \|X_\infty\|.$$

- Consequence: the Bayes regression function for the limit pair (X_∞, Y_∞):

$$f^*_{P_\infty}(x) := \mathbb{E}\left(Y_\infty \mid X_\infty = x\right) = \mathbb{E}\left(Y_\infty \mid \Theta_\infty = \theta(x)\right)$$

  - Depends on the angle only
  - Minimizes R_∞ under additional stability assumption (1)

Empirical tail risk minimization (Huet et al., 2023)

Input: $X_{1:n}, Y_{1:n}$ **i.i.d**

- Choose a class of prediction function $\mathcal{H}$, with $h : \mathbb{R}^d \to \mathbb{R}$
- Choose k
- Empirical Risk Minimization, **NO COMPLEXITY PENALTY**:

$$\underset{h \in \mathcal{H}}{\text{minimize}} \qquad \sum_{i \le k} (Y_{(i)} - h(\theta(X_{(i)})))^2$$

Output $\widehat{h}$

- Use $\widehat{h}(x)$ to predict Y if $\|x\|$ is large.

Excess risk guarantees: (Huet et al., 2023)

- Standard assumption that $\mathcal{H}$ is "VC subgraph", bounded + standard "pointwise measurability" assumptions

- Then, at finite level (quantile $1 - k/n$ of the norm)

$$\sup_{h \in \mathcal{H}} \left| \widehat{R}_k(h \circ \theta) - R_{t(n,k)}(h \circ \theta) \right| \leq_{w.\mathbb{P}.1-\delta} D_k = O\left(\sqrt{\frac{\log(1/\delta)}{k}} \right)$$

Also at infinity,

$$R_\infty(\widehat{f}_k) - R^*_\infty \leq D_k + B_1(k,n) + B_2(\mathcal{H})$$

Where $B_1(k,n)$ is sub-asymptotic bias $\to 0$ with additional total boundedness assumption on the class OR bounded angular density

Outline

Normalization trick (Clémençon and S., 2025), Proposition 4.1-a

- $(X, Z) \in \mathbb{R}^d \times \mathbb{R}$, where Z: target

- Assume (X, Z) multivariate heavy-tailed, $\mathbb{P}(X = 0) = 0$,

$$\mathbb{P}\left(r^{-1}(X, Z) \in \cdot \;\mid\; \|(X, Z)\| > r\right) \to \Pi_\infty(\cdot)$$

  $\Pi_\infty\{(x, z) : \|x\| > 1\} \neq 0$ [1]

- $Y = Z/\|X\|$.

- Then (X, Y) satisfies the 'one component RV' assumption.

[1] Z cannot be an order of magnitude larger than $\|X\|$

Normalization trick (Clémençon and S., 2025)

- If in addition: the density of (X, T) is is regularly varying, with some uniform convergence (Cai et al., 2011), and if the x-marginal of the limit density is lower bounded,

  Then also the regression function converges (in expectation) as in (1) and existing theory applies.

- This extends (simplifies) previous findings of Huet et al. (2023) who consider **nonlinear** transform $Z/\|(X, Z)\|$.

Outline

Outline

High dimensional issues

- Covariate dimension d large $\rightarrow$ typical class of predictors (e.g. linear predictors and variants) have complexity (VC dimension) growing above reasonable levels

- Of course, one should (and can easily) regularize the least-squares problem. This has been discussed at length in the high-dimnsional statistics literature.

- **Question:** Can we do this in our 'tail meta-algorithm', with some kind of guarantees? How difficult is it to adapt standard arguments (Hastie et al., 2015; Bühlmann and Van De Geer, 2011)? What kind of assumptions are necessary?

XLASSO: LASSO on extreme covariates

Section 5 in Clémençon and S. (2025).

- Choose a class of linear predictors

$$\mathcal{H} = \{h_\beta : x \mapsto \langle \beta, x \rangle,\ \beta \in \mathbb{R}^d\},$$

- This is a VC subgraph class of dimension $d + 1$

- Solve the angular LASSO problem on extremes:

$$\min_{\beta \in \mathbb{R}^d,} \frac{1}{2k} \sum_{i=1}^{k} (Y_{(i)} - h_\beta \circ \theta(X_{(i)}))^2 + \lambda \|\beta\|_1.$$

Related work (Aghbalou et al., 2024)

- 'Hard constrained LASSO', logistic regression (=binary classification)
- Focus: guarantees for risk estimation via Cross Validation
- Logistic loss (classification) $c(h_\beta(x, y)) = \log(1 + \exp(-\beta^\top \theta(x) y)$, $y \in \{-1, 1\}$
- Hard constrained ERM problem

$$\min_\beta \widehat{R}_k(h_\beta) \quad s.t. \ \|\beta\|_1 \leq C.$$

- Key differences:
    - Soft constraints (penalized loss) require more involved and LASSO-specific analysis
    - Binary target in Aghbalou et al. (2024) $\neq$ Regression

Asymptotic linear Model I

- Outside EVT, some modeling assumptions are typically required for LASSO guarantees (Hastie et al., 2015)

- It would be surprising not to need such assumptions in extrapolation settings

- Assumption: For some $\beta^* \in \mathbb{R}^d$,

$$Y = \langle \theta(X), \beta^* \rangle + b(X) + \sigma(x)\varepsilon,$$

where ε: unit noise, $|\sigma(x)| \leq M_\epsilon$ almost surely, and $b : \mathbb{R}^d \to \mathbb{R}$ is a bounded function that vanishes at infinity,

$$\bar{b}(t) := \sup_{x : r(x) > t} |b(x)| \xrightarrow[t \to \infty]{} 0.$$

- In this model: the required assumptions for regression on extremes Huet et al. (2023) are met.

Asymptotic linear Model II: unbounded target (informal version)

Linear heteroscedastic model with homogeneous variance:

$$S(x) = \|x\| s_\theta(\theta(x)),$$

X multivariate heavy tailed, and Target Z:

$$Z = \langle X, \beta^* \rangle + S(X)\varepsilon + o(\|X\|)$$

Rescaled target

$$Y = Z/(\|X\| \vee 1)$$

Asymptotic linear Model III: unbounded target (formal version)

X multivariate regularly varying, and

$$Z = X^\top \beta^* + B(X) + \sigma_z(X)\epsilon \quad ; \quad Y = Z/(\|X\| \vee 1),$$

where perturbation function B s.t.

- $\sup_{x \in \mathbb{R}^d} |B(x)|/\|x\| = M_B < \infty$
- $\sup_{\|x\|>t} |B(x)|/\|x\| \to 0$ [large noise allowed in the tail]
- $\sigma(x)/\|x\| \to \sigma_\theta(\theta(x))$ uniformly, $\sup_x \sigma(x)/(\|x\| \vee 1) = M_\varepsilon < \infty$

Outline

Notations, familiar objects (but in the tails)

- "Design matrix" of extreme angles

$$\mathbf{W} = (\theta(X_{(1)})^\top, \ldots, \theta(X_{(k)})^\top)^\top \in \mathbb{R}^{k \times p};$$

- Target $\mathbf{y} = (Y_{(1)}, \ldots, Y_{(k)}) \in \mathbb{R}^k$
- Residual vector $\mathbf{e} = \mathbf{y} - \mathbf{W}\beta^*$ (includes a bias term b).

Key Lemma

Prediction error, Bunea et al. (2007), Theorem 11.2-a in Hastie et al. (2015)

Assume $\lambda \geq 2k^{-1}\|\mathbf{W}^\top \mathbf{e}\|_\infty$. The (tail, in-sample) prediction error of the XLASSO estimator then satisfies

$$k^{-1}\|\mathbf{W}(\widehat{\beta} - \beta^*)\|_2^2 \leq 12\|\beta^*\|_1\lambda. \quad (2)$$

- Non random result: thus, it also applies to the top k order statistics
- "Prediction error?": $\mathbf{Z}\widehat{\beta} = \widehat{Y}$, and $(Z\beta^*)_i = f^*_\infty(X_{(i)})$

Key steps to next results

- Standard in lasso analysis: because $\lambda \geq 2k^{-1}\|\mathbf{W}^\top \mathbf{e}\|_\infty$ is necessary, one must control the latter 'empirical process' term to ensure that λ can be chosen small enough of order $1/\sqrt{k}$

- **W** is bounded! matrix of angular variables.

- Working conditionally to the $X_i's$, then integrating out, removes the problem of 'non-independence of top-k order statistics.

- With bounded noise (recall this is not so restrictive), McDiarmid inequalities works.

- Subgaussian noise would also be doable

XLASSO: Main result: minimal prediction guarantees

Theorem (XLASSO: prediction error guarantees)

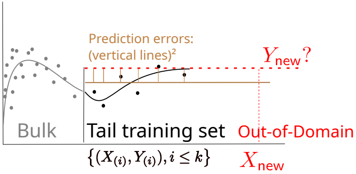
Let

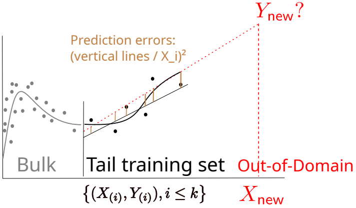
$$\lambda \geq M_\varepsilon \sqrt{\frac{\log(4d/\delta)}{2k}} + \bar{b}(t_{1-\tilde{k}(\delta/2)/n}),$$

where $\tilde{k}(\delta) \approx k$, $\tilde{k}(\delta) = k\left(1 + \sqrt{\frac{3\log(1/\delta)}{k}} + \frac{3\log(1/\delta)}{k}\right)$, *and* $\bar{b}(t) = \sup_{\|x\|>t} b(x)$.

Then w.p. at least $1 - \delta$,

$$k^{-1}\|\mathbf{W}^\top(\widehat{\beta} - \beta^*)\|_2^2 \leq 12\|\beta^*\|_1\lambda.$$

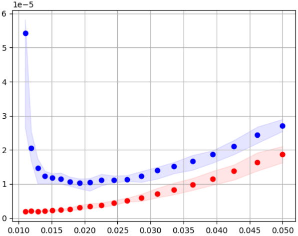




Experiments: Simulated data

- $Y = \langle \theta(X), \beta_0 \rangle + \frac{1}{\log(1+\|X\|)} \langle \theta(X), \beta_1 \rangle + \epsilon$,
- $\beta_1 \equiv 1$ and β_0 5-sparse, $d = 100$, $n = 10^4$, $d = 100$.
  $k \in [0.011n, 0.05n]$.

- Test set radial quantile: 0.99 ; 20 replications

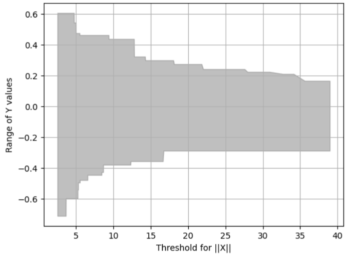


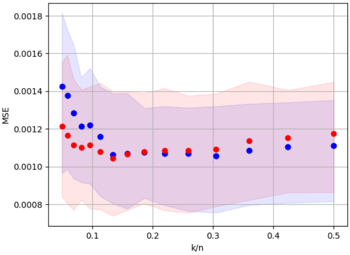
Red dots: XLASSO; Blue dots: linear regression,
Generalization prediction error on extreme test set:

$$n_{test}^{-1} \sum_{i=1}^{n_{test}} (Y_i - \hat{Y}_i)^2$$

Experiments: Real data

- **Industry Portfolio Dataset** (Meyer and Wintenberger, 2024; Huet et al., 2023). Target: $Z =$ "Transportation sector", $d = 49$, $n = 13577$.
- Target rescaling: $Y = Z/\|X\|$, X: other variables.
- Threshold for $\|X\|$: $1 - 0.005$ quantile for test, $1 - [0.05, 0.5]$ for train.
- left panel: boundedness of Y?





Conclusion

XLASSO: many more open questions than answer.

- One answer: Basic 'consistency result' (Bühlmann and Van De Geer, 2011): prediction error on the training data
- Out-of-sample (on new data point) prediction/estimation error?
- Estimation error control: conditions required about $\mathbf{W}^\top \mathbf{W}$ eigenvalues: realistic for extremes? Further work
- Just released

SOFTWARE

MLExtreme Python Package

`https://github.com/hi-paris/MLExtreme/`

  - Unsupervised: anomaly scoring with MV sets, support identification (feature clustering), PCA
  - Supervised: Classification, Regression (compatible with any learner with a fit and predict method, à la scikit-learn)
  - Tutorial notebooks

References I

Aghbalou, A., Bertail, P., Portier, F., and Sabourin, A. (2024). Cross-validation on extreme regions. *Extremes*, 27(4):505–555.

Bühlmann, P. and Van De Geer, S. (2011). *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media.

Bunea, F., Tsybakov, A., and Wegkamp, M. (2007). Sparsity oracle inequalities for the lasso. *Electronic Journal of Statistics*, 1:169–194.

Buriticá, G. and Engelke, S. (2024). Progression: an extrapolation principle for regression. *arXiv preprint arXiv:2410.23246*.

Cai, J., Einmahl, J., and De Haan, L. (2011). Estimation of extreme risk regions under multivariate regular variation. *The Annals of Statistics*, pages 1803–1826.

Clémençon, S., Jalalzai, H., Lhaut, S., Sabourin, A., and Segers, J. (2023). Concentration bounds for the empirical angular measure with statistical learning applications. *Bernoulli*, 29(4):2797–2827.

Clémençon, S. and S., A. (2025). Weak signals and heavy tails: Machine-learning meets extreme value theory. *arXiv preprint arXiv:2504.06984*.

de Carvalho, M., Ferrer, C., and Vallejos, R. (2025). A kolmogorov-arnold neural model for cascading extremes. *arXiv preprint arXiv:2505.13370*.

References II

de Carvalho, M., Kumukova, A., and Dos Reis, G. (2022). Regression-type analysis for multivariate extreme values. *Extremes*, 25(4):595–622.

De Haan, L. and Resnick, S. (1987). On regular variation of probability densities. *Stochastic processes and their applications*, 25:83–93.

Hastie, T., Tibshirani, R., and Wainwright, M. (2015). Statistical learning with sparsity. *Monographs on statistics and applied probability*, 143(143):8.

Huet, N., Clémençon, S., and Sabourin, A. (2023). On regression in extreme regions. *arXiv preprint arXiv:2303.03084*.

Jalalzai, H., Clémençon, S., and Sabourin, A. (2018). On binary classification in extreme regions. In *Advances in Neural Information Processing Systems*, pages 3092–3100.

Meyer, N. and Wintenberger, O. (2024). Multivariate sparse clustering for extremes. *Journal of the American Statistical Association*, 119(547):1911–1922.