Offre de stage - Internship Offer
Learning with heavy-tailed inputs: Out-of-domain Generalization on Extremes.

4-6-Month Internship leading to a PhD thesis - Spring 2025

Academic context This research internship takes place in the context of the ANR EXSTA project (EXtremes, STatistical learning and Applications (2024–2029) led by Anne Sabourin. The proposed internship is intended to lead to a doctoral thesis funded by the project. The potential PhD candidate will benefit from scientific interactions with other researchers in the field *e.g.* through workshops organised within the project's framework, in addition to usual participation in conferences.

Scientific Context Extreme Value Theory (EVT) is a field of probability and statistics concerned with tails of distributions, that is, regions of the sample space located far away from the bulk, associated with rare and extreme events. Let X be a random element (variable, vector, or function) of interest. One major goal of EVT is to provide probabilistic descriptions and statistical inference methods for the conditional distribution of X given large $\|X\|$, *i.e.* $\|X\| > t$, where $\|\cdot\|$ is a semi-norm and t is a large threshold (see *e.g.* the monographs De Haan and Ferreira (2006); Resnick (2008)). In applications, relevant thresholds t for probabilistic predictions may be as high as the largest observation among n realizations of X. Probabilistic extrapolation is then needed to use the information brought by a subsample of size $k_n \ll n$ composed of the observations with the largest semi-norms. This requires sound theoretical assumptions pertaining to the theory of regular variation and maximum domains of attraction, ensuring that a limit distribution $\mu = \lim \mathrm{Law}(t^{-1}X \,|\, \|X\| > t)$ exists as $t \to \infty$, up to suitable standardization. This stylized setting encompasses a wide range of applications in various disciplines where extremes have tremendous impact, such as climate science, insurance, environmental risks and industrial monitoring systems Beirlant et al. (2004).

In a supervised learning framework, training observations consist of pairs (X_i, Y_i) and the goal is to learn a good prediction function $f(X)$ to predict a new, unobserved Y. Machine learning and AI algorithms typically aim at minimizing an expected error $R(f) = E[\ell(f(X), Y)]$, for some loss function ℓ. In many contexts (covariate-shifts, climate change), extrapolation (or out-of-sample) properties of the predictors thus constructed are crucial, and obtaining good generalization properties on unobserved regions of the covariate space is key. Recently, there has been significant interest in the challenge of establishing guarantees for out-of-domain generalization (see *e.g.* Wang et al. (2022); Zhou et al. (2022) for a review of the ML literature on this topic) under specific assumptions.

Recent works focus on the problem of learning a predictor $\hat{f}_k$ based on a fraction k_n/n of the most extreme observations with guarantees regarding the risk on extreme regions $R_t(f) = E[\ell(f(X,Y)\|X\| > t]$, as $t \to \infty$. Existing works cover the problem of binary classification (Jalalzai et al. (2018, 2020); Clémençon et al. (2023)) and least squares regression (Huet et al. (2023)). In both contexts, generalization guarantees for the asymptotic risk as $t \to \infty$ have been obtained under natural regular variation assumptions on the pair (X, Y) (mainly, that $\mu_{XY} = \lim_{t\to\infty} \mathrm{Law}(t^{-1}X, Y \,|\, \|X\| > t)$ exists, while Y is bounded). A key common idea behind these works is to restrict the search of a good predictor to *angular* predictors of the kind $f(x) = f(\|x\|^{-1}x), x \neq 0$, which is shown to be legitimate under the aforementioned assumptions.

For simplicity, the theoretical study in both works is limited to Empirical Risk Minimization (ERM) algorithms *without* a penalty term. In addition, the regression problem analysed in Huet et al. (2023) covers least squares regression only. Also, the assumption that Y is bounded is made for simplicity only. With heavy-tailed targets, a natural situation in the context of extreme value analysis, non-linear transformations of the target are required in order to satisfy the boundedness assumptions.

Research Objectives The general purpose of this internship is to extend the scope of applications of the supervised learning methods described above to a wider class of learning algorithms. One main limitation of least squares regression is that the optimal predictor (*i.e.* the conditional expectation $E[Y|X]$) is not invariant under non-linear transformations of Y, indeed, in general, for such a transformation φ, $\varphi(E[Y|X]) \neq E|\varphi(Y)|X]$. As a starting point, the least-squares framework of Huet et al. (2023) will be extended to the quantile regression framework which, in contrast to the least squares setting, is compatible with non-linear

transformations. Indeed the median, say, of the transformed variable $\varphi(Y)$ is the same as the output of any monotone transformation φ applied to the median of Y. The first step will be to study the convergence of conditional quantiles of Y given X, as $\|X\| \to \infty$.

From a statistical learning perspective, we shall extend the ERM framework considered thus far to encompass penalized risk minimizations procedures amenable to high dimensional covariates or non-linear regression functions. SVM quantile regression (Takeuchi et al. (2006)) is a natural candidate for this purpose. The goal will be to obtain finite sample guarantees on the generalization error of quantile regression functions learnt with the k_n largest observations (w.r.t. the norm of X), and hopefully recover learning rates of comparable order as the ones obtained in the classical framework, with the full sample size n replaced with the reduced sample size k_n. The bottleneck is that these k_n largest observations may not be considered as an independent sample because they are order statistics of a full sample. However it is anticipated that proof techniques from recent works (Goix et al. (2015); Lhaut et al. (2022); Huet et al. (2023); Aghbalou et al. (2024)) based on conditioning arguments and concentration inequalities incorporating (small) variance terms can be leveraged for this purpose. Our first objective will be to obtain minimal guarantees (slow rates) associated with a control of Rademacher complexities, following Takeuchi et al. (2006). For an introduction to SVM's and proof techniques using Rademacher complexities of kernel classes, see Mohri et al. (2018) (Chapters 4,5,10).

On the longer term, the internship will serve as a preparation for the PhD thesis. Envisioned research projects include obtaining fast rates within the same SVM framework (Steinwart and Christmann (2011)), using sparsity inducing penalties (Zhang et al. (2016)), or exploring other learning approaches such as aggregation methods (*e.g.* random forests) or local predictors (k-NN) in similar supervised learning frameworks as described for this internship.

Supervisory team and contact The thesis will be hosted in the MAP5 laboratory, at Université Paris-Cité.

- Supervision (Internship + PhD thesis): Anne Sabourin (MAP5, Université Paris Cité, France), `anne.sabourin@math.cnrs.fr`, `https://helios2.mi.parisdescartes.fr/~asabouri/index.html#generalInfo`
- Co-supervision (PhD thesis): Clément Dombry (LMB, Université de Fanche-Comté, France), `clement.dombry@univ-fcomte.fr`, `https://cdombry.perso.math.cnrs.fr/`
- Envisioned collaboration with: Johan Segers (Department of Mathematics, KU Leuven, `jjjsegers@kuleuven.be`, `https://perso.uclouvain.be/johan.segers/`

The intern/PhD candidate will be offered travel opportunities in order to work with all three parties.

References

Aghbalou, A., Bertail, P., Portier, F., and Sabourin, A. (2024). Cross-validation on extreme regions. *Extremes*, pages 1–51.

Beirlant, J., Goegebeur, Y., Segers, J., and Teugels, J. L. (2004). *Statistics of Extremes: Theory and Applications*, volume 558. John Wiley & Sons.

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.

De Haan, L. and Ferreira, A. (2006). *Extreme Value Theory: An Introduction*. Springer.

Goix, N., Sabourin, A., and Clémençon, S. (2015). Learning the dependence structure of rare events: a non-asymptotic study. In *COLT Proceedings*, volume 40, pages 843–860. PMLR.

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 *NeurIPS Proceedings*, volume 31.

Jalalzai, H., Colombo, P., Clavel, C., Gaussier, E., Varni, G., Vignon, E., and Sabourin, A. (2020). Heavy-tailed representations, text polarity classification & data augmentation. In *NeurIPS Proceedings*, volume 33, pages 4295–4307.

Lhaut, S., Sabourin, A., and Segers, J. (2022). Uniform concentration bounds for frequencies of rare events. *Statist. Probab. Lett.*, 189:109610.

Mohri, M., Rostamizadeh, A., and Talwalkar, A. (2018). *Foundations of Machine Learning, second edition*. The MIT Press.

Resnick, S. I. (2008). *Extreme values, Regular Variation, and Point Processes*. Springer.

Steinwart, I. and Christmann, A. (2011). Estimating conditional quantiles with the help of the pinball loss. *Bernoulli*, 17(1):211–225.

Takeuchi, I., Le, Q. V., Sears, T. D., Smola, A. J., and Williams, C. (2006). Nonparametric quantile estimation. *Journal of machine learning research*, 7(7).

Wang, J., Lan, C., Liu, C., Ouyang, Y., Qin, T., Lu, W., Chen, Y., Zeng, W., and Philip, S. Y. (2022). Generalizing to unseen domains: A survey on domain generalization. *IEEE transactions on knowledge and data engineering*, 35(8):8052–8072.

Zhang, C., Liu, Y., and Wu, Y. (2016). On quantile regression in reproducing kernel hilbert spaces with the data sparsity constraint. *J. Mach. Learn. Res.*, 17(1):1374–1418.

Zhou, K., Liu, Z., Qiao, Y., Xiang, T., and Loy, C. C. (2022). Domain generalization: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4396–4415.