Cross Validation for rare events

Anass Aghbalou, François Portier, Patrice Bertail, Anne Sabourin

arXiv:2202.00488

GT-extremes, LPSM, February 2022

Outline

Motivation

- Cross-Validation (CV): widely used (even in extremes) for
  1. Model selection or Hyper-parameter selection
  2. Estimating the generalisation risk of a learning algorithm/Estimator.
- No existing theoretical work regarding **Cross-Validation for EVT algorithms/estimators**
- In many EVT works tuning parameters are present (let aside k)
- **Our wish:** Explore the question of **possible guarantees/pitfalls** for CV in an EVT/rare events context and make a **first theoretical step**.

Why is there a question? I

- *Just do Cross-Validation!*
- *Sure. However . . .*

1. For **model selection**: guarantees only in **specific settings**, *e.g*
   - density estimation Arlot 2008, Arlot& Lerasle 2008
   - LASSO Homrighausen and McDonald 2013; Xu et al. 2020.

   Selected parameter may be **inefficient/inconsistent**:
   - Bias for the K-fold
   - Dependence between folds in general

   see Arlot & Celisse 2010 or Wager 2020; Bates et al. 2021

Why is there a question? II

2. For **risk estimation**: Intuitively the CV estimate should be better (no over fitting issue) than the empirical risk on the training set.

Anyway:

- Hard to analyze (dependence between folds / bias).

- Most existing works: use *algorithmic stability* (changing one training point doesn't change much the output) Rogers and Wagner 1978; Kearns and Ron 1999; Bousquet and Elisseeff 2002

- Non-asymptotic guarantees: *Sanity-check* bounds Kearns and Ron 1999; Cornec 2009, 2017. **Of the same nature as the bounds on the empirical risk.**

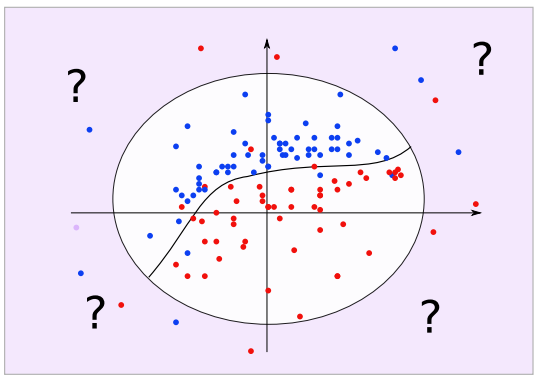
In this work we consider only purpose 2: risk estimation.

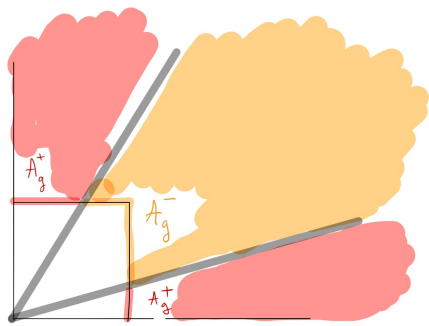
Outline

Statistical learning setting

- Classification problem with *i.i.d.* data $\mathcal{D}_n = Z_i = (X_i, Y_i), i \leq n$ in $\mathcal{Z}$; Collection $\mathcal{G}$ of classifiers; Cost function $c(g, Z)$

- Low probability region $\mathbb{A} = \{z = (x, y) : \ \|x\| > t_\alpha\} \subset \mathcal{Z}$ where t_α: quantile of the norm $\|\cdot\|$ (unknown). Framework formalized in Jalalzai et al. 2018





- (Generalization) risk of a candidate g on the rare region

$$\mathcal{R}_\alpha(g) = \mathbb{E}\left[c(g, Z) \mid Z \in \mathbb{A}\right].$$

Notations

- Ψ: a learning algorithmn, $\Psi(S) = \hat{g}_S$: output of Ψ trained on subsample $\{Z_i, i \in S\} \subset \mathcal{D}_n$

- **Goal**: Given $\mathcal{D}_n$, estimate the true risk of the output

$$\mathcal{R}_\alpha(\Psi([1:n]))$$

- Empirical risk of g evaluated on subsample indexed by $S \subset [1:n]$ of size n_S

$$\widehat{\mathcal{R}}_\alpha(g, S) = \frac{1}{\alpha n_S} \sum_{i \in S} c(g, Z_i)\ \mathbb{1}\{\|X_i\| > \|X_{(\lfloor \alpha n \rfloor)}\|\}.$$

Cross-Validation estimator:

Given a collection of Validation/Training sets
$V_j \subset \{1, \ldots, n\},\ T_j = \{1, \ldots n\} \setminus V_j,\ j = 1, \ldots, K$

$$\widehat{\mathcal{R}}_{\mathrm{CV},\alpha}(\Psi, V_{1:K}) = \frac{1}{K} \sum_{j=1}^{K} \widehat{\mathcal{R}}_\alpha(\Psi(T_j), V_j), \tag{1}$$

Our main purpose

Derive finite sample upper bounds (in probability) for the CV risk estimation error

$$|\mathrm{Error}_\alpha| = |\mathcal{R}_\alpha(\Psi([1:n])) - \widehat{\mathcal{R}}_{\mathrm{CV},\alpha}(\Psi, V_{1:K})|$$

Existing works, $\alpha = 1$, ERM setting (Ψ is a risk minimization rule) Cornec 2009, 2017 and Kearns & Ron 99:

- **K-fold:** with proba. $1-\delta$, $\quad |\mathrm{Error}_1| \leq \mathcal{O}\left(\sqrt{n^{-1}\log(1/\delta)}\right)$,
- **leave-one-out:** with proba. $1-\delta$, $\quad |\mathrm{Error}_1| \leq \mathcal{O}\left(\sqrt{n^{-1}}/\delta\right)$.

naive method: Divide by α each side of the inequality

$$\mathrm{Error}_\alpha \leq \mathcal{O}(\sqrt{n^{-1}}/\alpha) = \mathcal{O}(\sqrt{n}/k) \rightarrow \quad \textbf{FAILURE.}$$

(same as in existing { stat. learn. + EVT } works, Goix et al. 2015, etc. . .)

Assumptions

1. (ERM) The learning algorithm Ψ is based on Empirical Risk Minimization
$$\Psi(S) = \arg\min_{g \in \mathcal{G}} \widehat{\mathcal{R}}_\alpha(g, S).$$

2. (Exchangeable CV scheme) : for all $j \leq K$, $|V_j| = n_V$ for some $n_V \leq n$ and
$$\frac{1}{K} \sum_{j=1}^{K} \mathbb{1}\{\ell \in V_j\} = \frac{n_V}{n} \quad \forall \ell \in [1:n]$$

3. (Class complexity) collection $\mathcal{G}$ of classifiers has finite VC dimension $\mathcal{V}$.

4. (Bounded cost function) $\sup_{g,z} |c(g, z)| = 1$.

Outline

Large test sample sizes: exponential bound

Theorem 1 (ABPS 2022+)

Under assumptions 1-4, w.p. $1 - 15\delta$,

$$|\mathcal{R}_\alpha(\Psi([1:n])) - \widehat{\mathcal{R}}_{\mathrm{CV},\alpha}(\Psi, V_{1:K})| \leq$$
$$M\sqrt{\mathcal{V}}\left(\frac{1}{\sqrt{n_V \alpha}} + \frac{4}{\sqrt{n_T \alpha}}\right) + 20\sqrt{\frac{2\log(1/\delta)}{n\alpha}} +$$
$$\mathcal{O}\left(\frac{1}{n_T \alpha} + \frac{\log(1/\delta)}{n\alpha}\right)$$

for some universal constant $M > 0$.

- Both n_T and n_V must be 'large'.

Some remarks

- Constant M comes from chaining, Gine, Guillou, 2001, other constants not optimized
- M can be (probably) computed or replaced with $\log(n\alpha)$ term, as in Lhaut, Sabourin, Segers (21+) or CLémençon, Jalalzai, Lhaut, Sabourin Segers (22+)
- Matches Cornec 2017 for $\alpha = 1$ but improvement by factor $\sqrt{\alpha}$ over naive method. Different concentration tools accounting for low α.

Corollary: K-fold exponential bound

Corollary (ABPS 2022+)

For the $K-$*fold* with $K \geq 2$, under Assumptions 1-4, w.p. $1-\delta$,

$$|\mathcal{R}_\alpha(\Psi([1:n])) - \widehat{\mathcal{R}}_{\mathrm{CV},\alpha}(\Psi, V_{1:K})| \leq$$
$$5M\sqrt{\frac{\mathcal{V}K}{n\alpha}} + 20\sqrt{\frac{2\log(1/\delta)}{n\alpha}} +$$
$$\mathcal{O}\Big(\frac{1}{n\alpha} + \frac{\log(1/\delta)}{n\alpha}\Big)$$

for some universal constant $M > 0$.

Proof ideas: intermediate quantities (I)

Pseudo-empirical risk

$$\tilde{\mathcal{R}}_\alpha(g, S) = \frac{1}{\alpha n_S} \sum_{i \in S} c(g, O_i) \mathbb{1}\{\|X_i\| > t_\alpha\}.$$

Average pseudo-empirical risk

$$\tilde{\mathcal{R}}_{CV,\alpha}(\Psi, V_{1:K}) = \frac{1}{K} \sum_{j=1}^{K} \tilde{\mathcal{R}}_\alpha(\Psi(T_j), V_j).$$

- Deviation term stemming from $|X_{(\lfloor n\alpha \rfloor)} - t_\alpha|$:

$$\mathrm{D}_{\mathrm{t}_\alpha} = |\widehat{\mathcal{R}}_{\mathrm{CV},\alpha} - \tilde{\mathcal{R}}_{CV,\alpha}|(\Psi, V_{1:K})$$

$\rightarrow$ Bernstein inequality, low variance of $\mathbb{1}\{\|X_i\| > t_\alpha\}$.

Proof ideas: intermediate quantities (II)

Average 'true' risk of the trained rules $\left(\Psi(T_j)\right)_{0\leq j\leq K}$

$$\mathcal{R}_{CV,\alpha}(\Psi, V_{1:K}) = \frac{1}{K}\sum_{j=1}^{K} \mathcal{R}_\alpha(\Psi(T_j))$$

Recall the average pseudo empirical risk $\tilde{\mathcal{R}}_{CV,\alpha}(\Psi, V_{1:K})$ previous slide)

- Deviation term stemming from $|\tilde{\mathcal{R}}_\alpha(\Psi(T_j), V_j) - \mathcal{R}_\alpha(\Psi(T_j))|$:

$$\mathrm{D_{cv}} = |\tilde{\mathcal{R}}_{CV,\alpha} - \mathcal{R}_{CV,\alpha}(\Psi, V_{1:K})|$$

$\rightarrow$ tools: Uniform deviations of the empirical risk over $\mathcal{G}$, Bernstein-type bouded difference inequality McDiarmid 98 and Rademacher bounds for VC classes (Gine Guillou 2001), conditioning upon $\{\|X\| > t_{n_\alpha}\}$.
N.B: similarities with e.g. Goix et al. 2015 but specific quantity here

$$Z = \frac{1}{K}\sum_{j=1}^{K} \sup_{g\in\mathcal{G}} |\tilde{\mathcal{R}}_\alpha(g, V_i) - \mathcal{R}_\alpha(g)|$$

upper bound involves a term $\mathcal{O}(\sqrt{\mathcal{V}/n_V})$.

Proof ideas: intermediate quantities (III)

Recall the average true risk of the $\Psi(V_j)$'s , $\mathcal{R}_{CV,\alpha}(\Psi, V_{1:K})$.

- Biais term (comes from $|\mathcal{R}_\alpha(\Psi(V_j)) - \mathcal{R}_\alpha(\Psi([1:n]))|$

$$\text{Bias} = |\mathcal{R}_{CV,\alpha}(\Psi, V_{1:K}) - \mathcal{R}_\alpha\big(\Psi([1:n])\big)|$$

$\rightarrow$ Controlled using specific ERM nature of Ψ:

Both $\mathcal{R}_\alpha(\Psi(V_j))$ and $\mathcal{R}_\alpha(\Psi([1:n]))|$ are $\mathcal{O}\left(\sqrt{\mathcal{V}\log(1/\delta)/(\alpha n_T)}\right)$-close to $\mathcal{R}^*_\alpha = \inf_g \mathcal{R}_\alpha(g)$.

- Final step

$$\text{Error}_\alpha \leq \text{D}_{\text{t}_\alpha} + \text{D}_{\text{cv}} + \text{Bias}.$$

Small test sample sizes: Polynomial bound

Theorem 2 (ABPS 2022+)

Under assumptions 1-4, w.p. $1 - 18\delta$,

$$|\mathcal{R}_\alpha(\Psi([1:n])) - \widehat{\mathcal{R}}_{\mathrm{CV},\alpha}(\Psi, V_{1:K})| \leq 9M\sqrt{\frac{\mathcal{V}}{n_T\alpha}} + \frac{5M\sqrt{\mathcal{V}} + M'}{\delta\sqrt{n_T\alpha}} + \frac{9}{n_T\alpha}$$

for some universal constant' $M, M' > 0$.

- Corollary for the leave-p-out: idem with $n_T = n - p \sim n$.
- Matches Kearns & Ron 99, Cornec 2017 for $\alpha = 1$, similar improvement for $\alpha < 1$ as before.
- Tightness of rate $1/\delta$? Yes for $\alpha = 1$ (Kearns & Ron 99)

Proof ideas

- Cannot use the previous decomposition $\mathrm{Error}_\alpha \leq \mathrm{D}_{\mathrm{t}_\alpha} + \mathrm{D}_{\mathrm{cv}} + \mathrm{Bias}$ because D_{cv} involves $n_V = p$ small.

- Start similarly with $\mathrm{Error}_\alpha \leq \mathrm{Bias} + \left|\widehat{\mathcal{R}}_{\mathrm{CV},\alpha}(\Psi, V_{1:K}) - \mathcal{R}_{CV,\alpha}(\Psi, V_{1:K})\right|$ but bound differently the deviation term, in particular $|\tilde{\mathcal{R}}_{CV,\alpha} - \mathcal{R}_{CV,\alpha}|$.

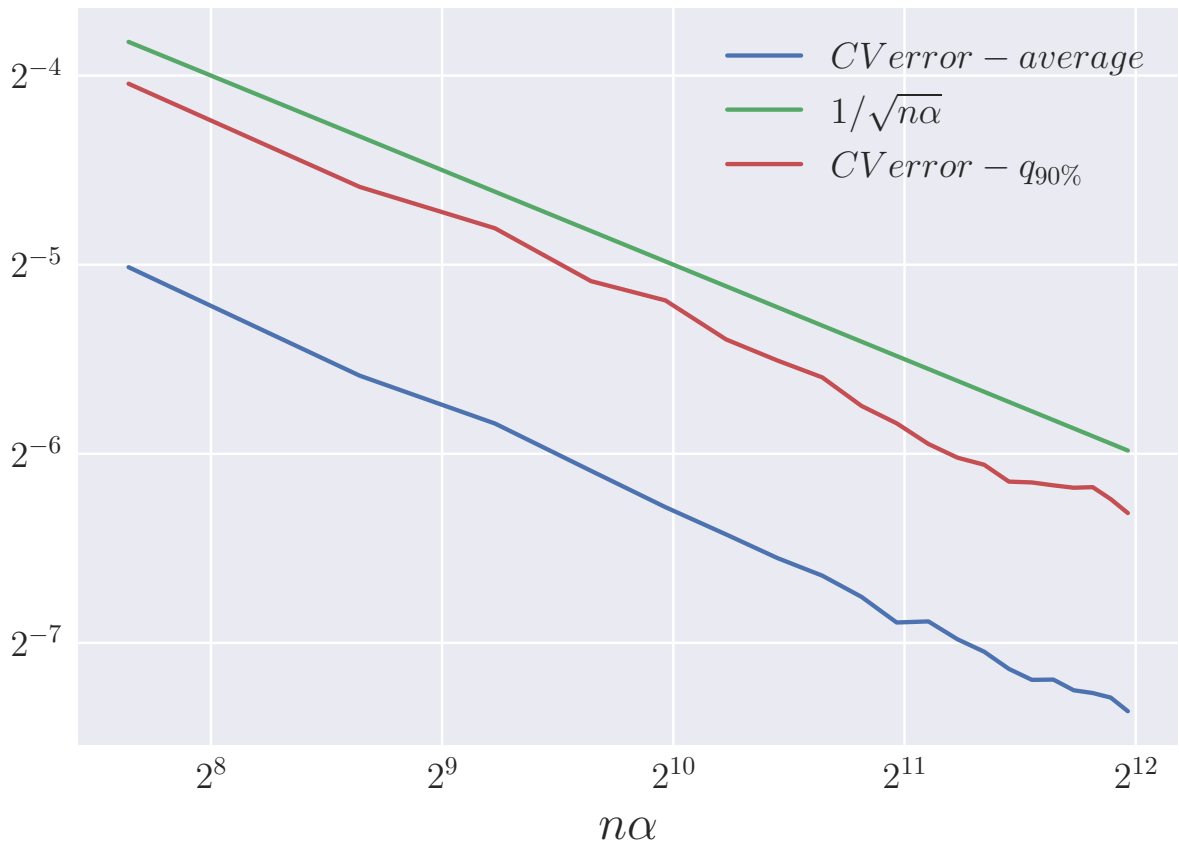
- Ingredient 1: $\mathbb{E}\left(\tilde{\mathcal{R}}_{CV,\alpha} - \mathcal{R}_{CV,\alpha}\right) = 0$
- Ingredient 2: $\widehat{\mathcal{R}}_{\mathrm{CV},\alpha}(\Psi, V_{1:K}) \geq \widehat{\mathcal{R}}_\alpha(\Psi([1:n]), [1:n])$ (see Kearns & Ron 99)
- Key step: Markov-type inequality

$$\mathbb{P}(\widehat{\mathcal{R}}_{\mathrm{CV},\alpha}(\Psi, V_{1:K}) - \mathcal{R}_{CV,\alpha}(\Psi, V_{1:K}) \geq t) \leq \frac{\mathbb{E}\left(\mathrm{D}_{\mathrm{t}_\alpha} + \mathrm{Bias} + \left|\hat{\mathcal{R}}_\alpha(\Psi_\alpha([1:n]), [1:n]) - \mathcal{R}_\alpha(\Psi_\alpha([1:n]))\right|\right)}{t}$$

Outline

Error rate $1/\sqrt{n\alpha}$?

- Toy example: simulated data, dimension 1,
  Class distributions: student, threshold classifier, Hamming loss
- $n = 2.10^4$, $\alpha \in [1\%, 20\%]$
- Average absolute error of the K-fold ($K = 10$) and upper quantile at level 0.90, logarithmic scale, over 10^4 experiments.



Discussion, perspectives

- Replacing ERM assumption with algorithmic stability $\rightarrow$ wider class of algorithms and improved bounds for the l-p-o.
- Extension to other rare events (imbalanced classification)?
- Beyond sanity check bounds? (even for $\alpha = 1$?)
- Model/parameter selection?
  - Effective choice (with good generalization risk) with finite collection of parameters: consequence of our results ($\log |M|$ multiplicative term)
  - Selecting the right model: much harder in general.

Bibliography I

- **A. Aghbalou, P. Bertail, F. Portier, A. Sabourin (2022) Cross-Validation for Rare Events. ArXiv 2202.0488v1.**
- S. Arlot (2008). V-fold cross-validation improved: V-fold penalization. URL https://hal.archives-ouvertes.fr/ hal-00239182.
- S. Arlot and M. Lerasle (2008). Choice of v for v-fold cross-validation in least-squares density estimation. Journal of Machine Learning Research, 17(208):1–50, 2016.
- S. Arlot and A. Celisse (2010). A survey of cross-validation procedures for model selection. Statistics surveys, 4:40–79.
- S. Bates, T. Hastie, and R. Tibshirani (2021). Cross-validation: what does it estimate and how well does it do it? arXiv preprint arXiv:2104.00673.
- S. Clémençon, H. Jalalzai, S. Lhaut, A. Sabourin, and J. Segers (2022). Concentration bounds for the empirical angular measure with statistical learning applications, arXiv:2104.03966v2
- M. Cornec (2009). Probability bounds for the cross-validation estimate in the context of the statisticallearning theory and statistical models applied to economics and finance. Thesis, Université de Paris-Nanterre.

Bibliography II

- **M. Cornec (2017). Concentration inequalities of the cross-validation estimator for empirical risk minimizer. Statistics, 51(1):43–60.**
- E. Giné and A. Guillou (2001). On consistency of kernel density estimators for randomly cen- sored data: rates holding uniformly over adaptive intervals. Annales de l'IHP Probabilités et statistiques, 37(4):503–522.
- **N. Goix, A. Sabourin, and S. Clémençon (2015). Learning the dependence structure of rare events: a non-asymptotic study. In Conference on Learning Theory, pages 843–860.**
- D. Homrighausen nad D. McDonald (2013). The lasso, persistence, and cross-validation. In ICML, pp 1031–1039. PMLR, .
- **H. Jalalzai, S. Clémençon, and A. Sabourin (2018). On binary classification in extreme re- gions. In NeurIPS, pages 3096–3104.**
- **M. Kearns and D. Ron (1999). Algorithmic stability and sanity-check bounds for leave-one-out** cross-validation. Neural computation, 11(6):1427–1453.
- S. Lhaut, A. Sabourin, and J. Segers (2021). Uniform concentration bounds for frequencies of rare events. arXiv preprint arXiv:2110.05826.

Bibliography III

- C McDiarmid (1998). Concentration. Probabilistic Methods for Algorithmic Discrete Mathematics, pages 195–248.
- S. Wager (2020). Cross-validation, risk estimation, and model selection: Comment on a paper by rosset and tibshirani. Journal of the American Statistical Association, 115(529):157–160.
- N. Xu, T. Fisher, and J. Hong (2020). Rademacher upper bounds for cross-validation errors with an application to the lasso. arXiv preprint arXiv:2007.15598.