# Generalized test utilities for long-tail performance in extreme multi-label classification

Erik Schultheis[1] Marek Wydmuch[2] Wojtek Kotłowski[2]
Rohit Babbar[1,3] Krzysztof Dembczyński[2,4]

[1]Aalto University, Helsinki, Finland [2]Poznan University of Technology, Poland
[3]University of Bath,UK [4]Yahoo! Research, New York, USA

NeurIPS '23, December 10–16, 2023, New Orleans, USA

# Extreme multi-label classification (XMLC)

## Multi-label classification:

$$\boldsymbol{x} \in \mathcal{X} \rightarrow \boldsymbol{y} \in \mathcal{Y} := \{0, 1\}^m$$

e.g.:

| | $y_1$ | $y_2$ | $y_3$ | $\cdots$ | $y_m$ |
|---|---|---|---|---|---|
| $\boldsymbol{y} =$ | 0 | 1 | 1 | ... | 0 |

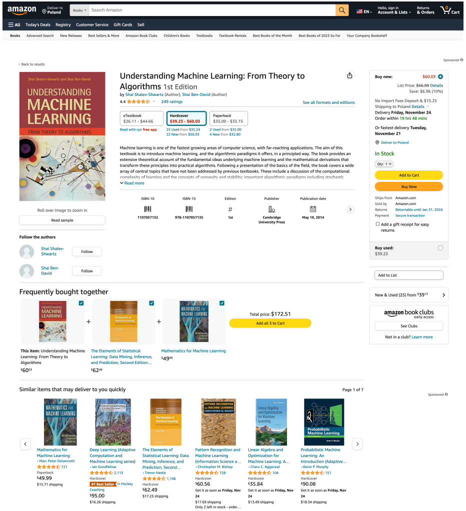
# Extreme multi-label classification (XMLC)

## Multi-label classification:

$$\boldsymbol{x} \in \mathcal{X} \to \boldsymbol{y} \in \mathcal{Y} := \{0,1\}^m$$

## Extreme multi-label classification:

- a **large** number of **labels** $m$ ($\geq 10^5$),
- a label vector $\boldsymbol{y}$ is **very sparse**, $\|\boldsymbol{y}\|_1 \ll m$,
- many problems are naturally **budgeted at** $k$ (requirement for a prediction $\hat{\boldsymbol{y}}$: $\|\hat{\boldsymbol{y}}\|_1 = k$),



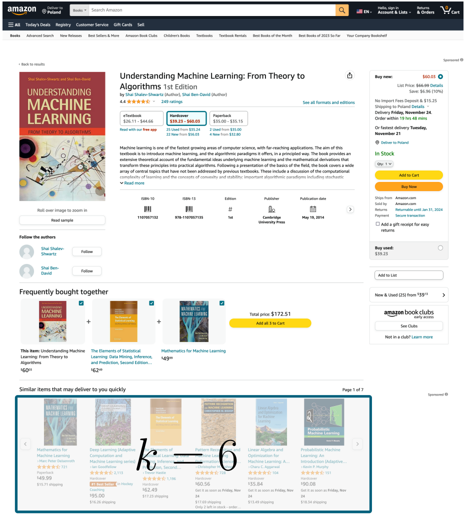
# Extreme multi-label classification (XMLC)

## Multi-label classification:

$$\boldsymbol{x} \in \mathcal{X} \rightarrow \boldsymbol{y} \in \mathcal{Y} := \{0,1\}^m$$

## Extreme multi-label classification:

- a **large** number of **labels** $m$ ($\geq 10^5$),
- a label vector $\boldsymbol{y}$ is **very sparse**, $\|\boldsymbol{y}\|_1 \ll m$,
- many problems are naturally **budgeted at** $k$ (requirement for a prediction $\hat{\boldsymbol{y}}$: $\|\hat{\boldsymbol{y}}\|_1 = k$),



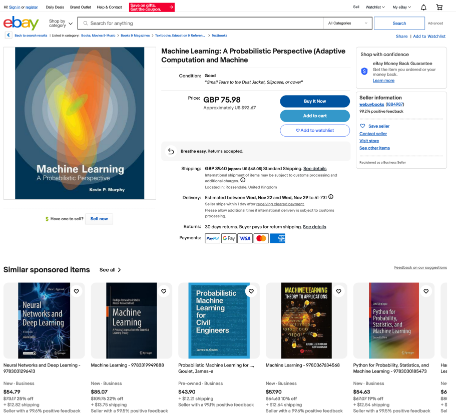
# Extreme multi-label classification (XMLC)

## Multi-label classification:

$$\boldsymbol{x} \in \mathcal{X} \rightarrow \boldsymbol{y} \in \mathcal{Y} := \{0, 1\}^m$$

## Extreme multi-label classification:

- a large number of labels $m$ ($\geq 10^5$),
- a label vector $\boldsymbol{y}$ is very sparse, $\|\boldsymbol{y}\|_1 \ll m$,
- many problems are naturally budgeted at $k$ (requirement for a prediction $\hat{\boldsymbol{y}}$: $\|\hat{\boldsymbol{y}}\|_1 = k$),



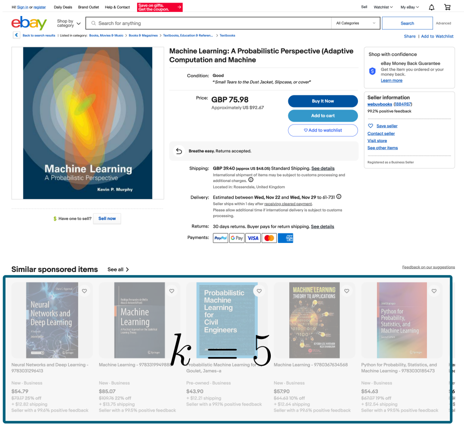
# Extreme multi-label classification (XMLC)

## Multi-label classification:

$$\boldsymbol{x} \in \mathcal{X} \rightarrow \boldsymbol{y} \in \mathcal{Y} := \{0,1\}^m$$

## Extreme multi-label classification:

- a **large** number of **labels** $m$ ($\geq 10^5$),
- a label vector $\boldsymbol{y}$ is **very sparse**, $\|\boldsymbol{y}\|_1 \ll m$,
- many problems are naturally **budgeted at** $k$ (requirement for a prediction $\hat{\boldsymbol{y}}$: $\|\hat{\boldsymbol{y}}\|_1 = k$),




$k = 5$

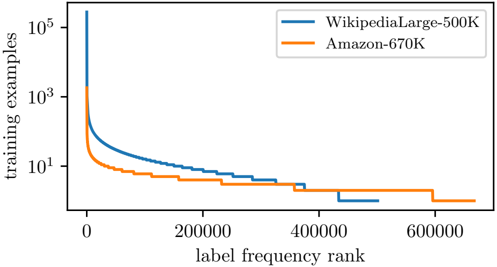
# Extreme multi-label classification (XMLC)

## Multi-label classification:

$$\boldsymbol{x} \in \mathcal{X} \to \boldsymbol{y} \in \mathcal{Y} := \{0, 1\}^m$$

## Extreme multi-label classification:

- a **large** number of **labels** $m$ ($\geq 10^5$),
- a label vector $\boldsymbol{y}$ is **very sparse**, $\|\boldsymbol{y}\|_1 \ll m$,
- many problems are naturally **budgeted at $k$** (requirement for a prediction $\hat{\boldsymbol{y}}$: $\|\hat{\boldsymbol{y}}\|_1 = k$),
- **long-tail distribution** of labels.



# Problem with long tail and common metrics budgeted at $k$

## Standard instance-wise metrics, e.g.:

$$\text{Precision@}k(\boldsymbol{Y}, \hat{\boldsymbol{Y}}) := \frac{1}{n} \sum_{i=1}^{n} \frac{1}{k} \sum_{j=1}^{m} y_{ij} \hat{y}_{ij}$$

## Problem with long tail and common metrics budgeted at $k$

### Standard instance-wise metrics, e.g.:

$$\text{Precision@}k(\boldsymbol{Y}, \hat{\boldsymbol{Y}}) := \frac{1}{n} \sum_{i=1}^{n} \frac{1}{k} \sum_{j=1}^{m} y_{ij} \hat{y}_{ij}$$

Table: Performance measures (%) on AmazonCat-13k of a classifier trained on the full set of labels and a classifier trained with only 1k head (most frequent) labels.

| Metric | full labels | | | head labels | | |
|---|---|---|---|---|---|---|
| | @1 | @3 | @5 | @1 (diff.) | @3 (diff.) | @5 (diff.) |
| Precision | 93.03 | 78.51 | 63.74 | 93.08 (+0.05%) | 76.42 (-2.66%) | 58.21 (-8.67%) |
| nDCG | 93.03 | 87.25 | 85.35 | 93.08 (+0.05%) | 85.75 (-1.71%) | 80.91 (-5.19%) |
| PS-Precision | 49.76 | 62.63 | 70.35 | 49.07 (-1.39%) | 57.71 (-7.84%) | 57.41 (-18.40%) |

# Problem with long tail and common metrics budgeted at $k$

## Standard instance-wise metrics, e.g.:

$$\text{Precision@}k(\boldsymbol{Y}, \hat{\boldsymbol{Y}}) := \frac{1}{n} \sum_{i=1}^{n} \frac{1}{k} \sum_{j=1}^{m} y_{ij} \hat{y}_{ij}$$

## Metrics that linearly decompose over labels:

$$\Psi@k(\boldsymbol{Y}, \hat{\boldsymbol{Y}}) = \sum_{j=1}^{m} \psi^j(\boldsymbol{y}_{:j}, \hat{\boldsymbol{y}}_{:j})$$

Table: Performance measures (%) on AmazonCat-13k of a classifier trained on the full set of labels and a classifier trained with only 1k head (most frequent) labels.

| Metric | full labels | | | head labels | | |
|---|---|---|---|---|---|---|
| | @1 | @3 | @5 | @1 (diff.) | @3 (diff.) | @5 (diff.) |
| Precision | 93.03 | 78.51 | 63.74 | 93.08 (+0.05%) | 76.42 (-2.66%) | 58.21 (-8.67%) |
| nDCG | 93.03 | 87.25 | 85.35 | 93.08 (+0.05%) | 85.75 (-1.71%) | 80.91 (-5.19%) |
| PS-Precision | 49.76 | 62.63 | 70.35 | 49.07 (-1.39%) | 57.71 (-7.84%) | 57.41 (-18.40%) |

# Problem with long tail and common metrics budgeted at $k$

## Standard instance-wise metrics, e.g.:

$$\text{Precision}@k(\boldsymbol{Y}, \hat{\boldsymbol{Y}}) := \frac{1}{n}\sum_{i=1}^{n}\frac{1}{k}\sum_{j=1}^{m} y_{ij}\hat{y}_{ij}$$

## Metrics that linearly decompose over labels:

$$\Psi@k(\boldsymbol{Y}, \hat{\boldsymbol{Y}}) = \sum_{j=1}^{m}\psi^{j}(\boldsymbol{y}_{:j}, \hat{\boldsymbol{y}}_{:j})$$

Table: Performance measures (%) on AmazonCat-13k of a classifier trained on the full set of labels and a classifier trained with only 1k head (most frequent) labels.

| Metric | full labels | | | head labels | | |
|---|---|---|---|---|---|---|
| | @1 | @3 | @5 | @1 (diff.) | @3 (diff.) | @5 (diff.) |
| Precision | 93.03 | 78.51 | 63.74 | 93.08 (+0.05%) | 76.42 (-2.66%) | 58.21 (-8.67%) |
| nDCG | 93.03 | 87.25 | 85.35 | 93.08 (+0.05%) | 85.75 (-1.71%) | 80.91 (-5.19%) |
| PS-Precision | 49.76 | 62.63 | 70.35 | 49.07 (-1.39%) | 57.71 (-7.84%) | 57.41 (-18.40%) |
| Macro-Precision | 13.28 | 32.65 | 44.16 | 4.31 (-67.54%) | 5.28 (-83.82%) | 4.32 (-90.21%) |
| Macro-Recall | 1.38 | 11.06 | 30.57 | 0.47 (-65.61%) | 2.69 (-75.71%) | 4.10 (-86.59%) |
| Macro-F1 | 2.26 | 14.67 | 32.84 | 0.74 (-67.37%) | 3.10 (-78.88%) | 3.77 (-88.51%) |
| Coverage | 15.19 | 40.53 | 60.88 | 5.11 (-66.32%) | 7.37 (-81.82%) | 7.52 (-87.65%) |

# Our contributions

- We analyze the problem of **optimization** of general family of metrics linearly decomposeable over labels calculated at $k$ under **expected test utility framework (ETU)**

$$\Psi@k(\boldsymbol{Y}, \hat{\boldsymbol{Y}}) = \sum_{j=1}^{m} \psi^j(\boldsymbol{y}_{:j}, \hat{\boldsymbol{y}}_{:j}) \,, \quad \hat{\boldsymbol{Y}}^\star = \operatorname*{argmax}_{\hat{\boldsymbol{Y}} \in \mathcal{Y}_k^n} \mathbb{E}_{\boldsymbol{Y}|\boldsymbol{X}} [\Psi@k(\boldsymbol{Y}, \hat{\boldsymbol{Y}})] \,.$$

# Our contributions

- We analyze the problem of **optimization** of general family of metrics linearly decomposeable over labels calculated at $k$ under **expected test utility framework (ETU)**

$$\Psi@k(\boldsymbol{Y}, \hat{\boldsymbol{Y}}) = \sum_{j=1}^{m} \psi^j(\boldsymbol{y}_{:j}, \hat{\boldsymbol{y}}_{:j}) , \quad \hat{\boldsymbol{Y}}^\star = \operatorname*{argmax}_{\hat{\boldsymbol{Y}} \in \mathcal{Y}_k^n} \mathbb{E}_{\boldsymbol{Y}|\boldsymbol{X}}[\Psi@k(\boldsymbol{Y}, \hat{\boldsymbol{Y}})] .$$

- Our framework only **requires** the probability estimates of individual labels for each instance $\boldsymbol{\eta}(\boldsymbol{x}) = (\eta_1(\boldsymbol{x}), \ldots, \eta_m(\boldsymbol{x})) := \mathbb{E}_{\boldsymbol{y}|\boldsymbol{x}}[\boldsymbol{y}] \rightarrow$ **easy to apply** on-top of existing classifiers

# Our contributions

- We analyze the problem of **optimization** of general family of metrics linearly decomposeable over labels calculated at $k$ under **expected test utility framework (ETU)**

$$\Psi@k(\boldsymbol{Y}, \hat{\boldsymbol{Y}}) = \sum_{j=1}^{m} \psi^j(\boldsymbol{y}_{:j}, \hat{\boldsymbol{y}}_{:j}) \,, \quad \hat{\boldsymbol{Y}}^\star = \operatorname*{argmax}_{\hat{\boldsymbol{Y}} \in \mathcal{Y}_k^n} \mathbb{E}_{\boldsymbol{Y}|\boldsymbol{X}}[\Psi@k(\boldsymbol{Y}, \hat{\boldsymbol{Y}})] \,.$$

- Our framework only **requires** the probability estimates of individual labels for each instance $\boldsymbol{\eta}(\boldsymbol{x}) = (\eta_1(\boldsymbol{x}), \ldots, \eta_m(\boldsymbol{x})) \coloneqq \mathbb{E}_{\boldsymbol{y}|\boldsymbol{x}}[\boldsymbol{y}] \rightarrow$ **easy to apply** on-top of existing classifiers
- We provide:
  - optimal prediction rules,
  - efficient approximations with guarantees,
  - regret bounds quantifying influence of label probability estimation error,
  - general algorithm, based on block coordinate ascent, that scales to XMLC problems.

# Thank you for your attention

Poster: Thursday, December 14, Poster Session 5, #1025

Paper: `https://arxiv.org/pdf/2311.05081.pdf`