

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

Erik Schultheis¹ Marek Wydmuch² Wojtek Kotłowski²
Rohit Babbar^{1,3} Krzysztof Dembczyński^{2,4}

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



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

Extreme multi-label classification (XMLC)

Multi-label classification:

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

e.g.:

	y_1	y_2	y_3	\dots	y_m
$\mathbf{y} =$	0	1	1	\dots	0

Extreme multi-label classification (XMLC)

Multi-label classification:

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

Extreme multi-label classification:

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

The screenshot shows the Amazon product page for the book "Understanding Machine Learning: From Theory to Algorithms, 1st Edition" by Shai Shalev-Shwartz and Shai Ben-David. The page includes the book cover, a star rating of 4.4, and various purchase options (hardcover, paperback, Kindle). The description highlights the book's focus on the theoretical foundations of machine learning. Below the product details, there are sections for "Follow the authors", "Frequently bought together" (showing a bundle of three books for \$172.51), and "Similar items that may deliver to you quickly" (a list of related machine learning books).

Extreme multi-label classification (XMLC)

Multi-label classification:

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

Extreme multi-label classification:

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

The screenshot shows an eBay product page for the book "Machine Learning: A Probabilistic Perspective (Adaptive Computation and Machine)". The price is GBP 75.98 (approximately US \$92.47). The page includes a "Buy it Now" button, an "Add to cart" button, and an "Add to wish list" button. Below the main product information, there are sections for shipping, delivery, and returns. At the bottom, there is a "Similar sponsored items" section displaying five other books for sale, each with its title, price, and shipping information.

Item	Price	Shipping
Neural Networks and Deep Learning	\$64.79	\$10.27 shipping
Machine Learning	\$66.07	\$10.75 shipping
Probabilistic Machine Learning for ...	\$43.90	\$12.21 shipping
Machine Learning: THEORY TO APPLICATIONS	\$67.90	\$10.44 shipping
Python for Probability, Statistics, and Machine Learning	\$64.43	\$12.04 shipping

Extreme multi-label classification (XMLC)

Multi-label classification:

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

Extreme multi-label classification:

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

The screenshot shows an eBay listing for the book "Machine Learning: A Probabilistic Perspective (Adaptive Computation and Machine)" by Kevin P. Murphy. The price is GBP 75.98. Below the main listing, there is a section titled "Similar sponsored items" which contains five book listings. A red box highlights this section, and a red "k=5" is written over it, indicating a budget constraint on the number of items shown.

Item	Price	Shipping
Neural Networks and Deep Learning	\$54.79	\$10.00 shipping
Machine Learning	\$86.07	\$10.75 shipping
Probabilistic Machine Learning for Civil Engineers	\$43.90	\$10.00 shipping
Machine Learning: Theory and Applications	\$67.90	\$10.00 shipping
Pattern Recognition, Statistics, and Machine Learning	\$54.43	\$10.00 shipping

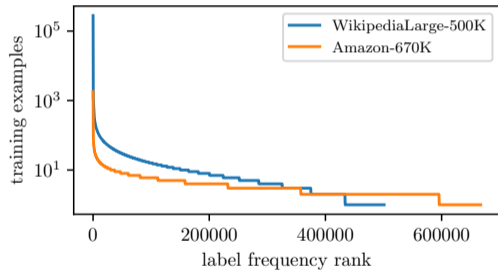
Extreme multi-label classification (XMLC)

Multi-label classification:

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

Extreme multi-label classification:

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



Problem with long tail and common metrics budgeted at k

Standard instance-wise metrics, e.g.:

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

Problem with long tail and common metrics budgeted at k

Standard instance-wise metrics, e.g.:

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

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.:

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

Metrics that linearly decompose over labels:

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

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.:

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

Metrics that linearly decompose over labels:

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

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 decomposable over labels calculated at k under **expected test utility framework (ETU)**

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

Our contributions

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

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

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

Our contributions

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

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

- Our framework only **requires** the probability estimates of individual labels for each instance $\boldsymbol{\eta}(\mathbf{x}) = (\eta_1(\mathbf{x}), \dots, \eta_m(\mathbf{x})) := \mathbb{E}_{\mathbf{y}|\mathbf{x}}[\mathbf{y}] \rightarrow$ **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>