Unbiased Estimates for Multilabel Reductions of Extreme Classification with Missing Labels

Anonymous Author(s)
Affiliation
Address
email

Abstract

This paper considers the missing-labels problem in the extreme multilabel classification (XMC) setting, i.e. a setting with a very large label space. The goal in XMC often is to maximize either precision or recall of the top-ranked predictions, which can be achieved by reducing the multilabel problem into a series of binary (One-vs-All) or multiclass (Pick-all-Labels) problems. Missing labels are a ubiquitous phenomenon in XMC tasks, yet the interaction between missing labels and multilabel reductions has hitherto only been investigated for the case of One-vs-All reduction. In this paper, we close this gap by providing unbiased estimates for general (non-decomposable) multilabel losses, which enables unbiased estimates of the Pick-all-Labels reduction, as well as the normalized reductions which are required for consistency with the recall metric. We show that these estimators suffer from increased variance and may lead to ill-posed optimization problems. To address this issue, we propose to use convex upper bounds which trade off an increase in bias against a strong decrease in variance.

1 Introduction

Extreme multilabel classification (XMC) is a machine learning setting in which the goal is to predict a small subset of positive (or relevant) labels for each data instance out of a very large (thousands to millions) set of possible labels. Such problems arise for example when annotating large encyclopedia [7, 30], in fine-grained image classification [9], and next-word prediction [26]. Further applications of XMC are recommendation systems, web-advertising and prediction of related searches [1, 31, 17, 6].

Typical datasets in these scenarios are very large, resulting in possibly billions of (data, label) pairs [4], making it impossible for human annotators to check each pair. Even annotating only a few samples fully in order to generate a clean test set can be prohibitively expensive. Therefore, both the available training- and test-data are likely to contain some errors. Fortunately, in many cases it is possible to constrain the structure of the labeling errors. Consider, for example, the case of tagging documents: Here, we can assume that each label with which the document has been tagged has been deemed relevant by the annotator, and thus is relatively surely a correct label. On the other hand, the annotator cannot possibly check hundreds of thousands of negative labels. This leads to the setting of missing labels investigated in this paper, in which only positive labels are affected by noise (they can go missing), whereas negative labels remain unchanged (no spurious labels). This model has been introduced to the XMC setting by Jain et al. [16], along with estimates for the *propensities*, the chance of a relevant label to be observed. Similar models are using in learning-to-rank[20, 29, 41] and recommendation systems[34, 14, 15]. For a formal definition of the setting we refer the reader to section 3, and for a more thorough discussion of prior works on missing labels and related settings to section 6.

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

A common strategy for learning XMC classifiers is to reduce the multilabel problem [38] into a series of binary [8, 3, 44] or multiclass [18, 42, 33] problems, which then can be solved using existing techniques. Such *loss reductions* can be shown to be consistent for the tasks of maximizing precision at k or recall at k, but never both at the same time [25]. For one of these methods, One-vs-All, adaptation to the missing labels setting has been shown to yield an improvement in propensity-scored precision (an unbiased estimate of precision@k) metrics [32]. The reductions consistent for precision lead to loss functions that can be decomposed into a sum of contributions from each label, which means the results of Natarajan et al. [28] can be applied. In contrast, the reductions consistent for recall contain a normalization term that is the inverse of the total number of true labels. This term is also necessary for calculating the recall metric itself, demonstrating the need for unbiased estimates for true, non-decomposable multilabel loss functions.

Contributions Our contributions are **1)** A mathematical model of the missing labels setting that describes the observed labels as a product of an (unknown) mask variable with the true labels. Crucially, this mask can be chosen to be *independent* of the labels (Theorem 1), enabling simple proofs for our theorems. **2)** The unique unbiased estimate (Theorems 2, 3) for arbitrary multilabel losses, and in particular for the loss functions arising from multilabel reductions. The unbiased estimate of a lower-bounded loss need not be lower-bounded, and even for bounded losses the unbiased estimate leads to an increase in variance. Therefore, we develop **3)** a convex upper-bound (Theorem 4) for losses based on the normalized Pick-all-Labels reduction. In the missing-labels setting, the generalization error is composed of two contributions: the error due to overfitting to the specific, observed noise-pattern, and the error because only a finite sample has been observed. We present empirical evidence **4)** that the former can be much stronger than the latter, and may be reduced by switching to the upper bounds.

In the main paper, we provide shortened proofs that illustrate the key steps. Detailed step-by-step proofs can be found in the appendix.

Notation Random variables will be denoted by capital letters $X, Y, \ldots$, whereas calligraphic letters denote sets and lower case letters their elements, $x \in \mathcal{X}, \ldots$. Vectors will be denoted by bold font, $\mathbf{y} \in \mathcal{Y}$, if we plan to make use of the fact that they can be decomposed into components $y_1, \ldots, y_k$, with $\mathbf{y}_{\neg k}$ denoting the vector of all components except the k'th. The letters f, g, h and ℓ are reserved for functions, i, j, k denote integers, $[k]$ is the set $\{1, \ldots, k\}$. We denote with $\mathcal{X}$ the *data space*, $\mathcal{Y} = \{0, 1\}^l$ the *label space* and $\hat{\mathcal{Y}} = \mathbb{R}^l$ the *prediction space*. A dataset is defined through the three random variables $X \in \mathcal{X}$, $\mathbf{Y} \in \mathcal{Y}$, and $\mathbf{Y}^* \in \mathcal{Y}$, that represent the *data*, *observed label*, and *ground truth label*. We mark quantities pertaining to the unobservable ground-truth with a superscript star and call $(X, \mathbf{Y}^*)$ the *clean data*.

2 Multilabel Reductions

In Menon et al. [25], five different reductions for turning the multilabel learning problem into a sum of binary or multiclass problems are presented (cf. appendix). In the following, let $\ell_{\mathrm{BC}} : \{0, 1\} \times \mathbb{R} \longrightarrow \mathbb{R}$ be a binary loss and $\ell_{\mathrm{MC}} : [l] \times \mathbb{R}^l \longrightarrow \mathbb{R}$ be a multiclass loss. Below, we present four of those reductions, and rearrange their loss functions so that a common pattern emerges.

For *one-vs-all* (OVA) reduction, each label is considered independently, meaning that for each instance l binary problems are to be solved. This leads to a loss function

$$\ell^*_{\mathrm{OvA}}(\mathbf{y}^*, \hat{\mathbf{y}}) = \sum_{j=1}^{l} \ell_{\mathrm{BC}}(y^*_j, \hat{y}_j) = \sum_{j=1}^{l} y^*_j \left(\ell_{\mathrm{BC}}(1, \hat{y}_j) - \ell_{\mathrm{BC}}(0, \hat{y}_j)\right) + \ell_{\mathrm{BC}}(0, \hat{y}_j). \tag{1}$$

In contrast, *pick-all-labels* (PAL) considers all the positive labels for each instance and tries to minimize their corresponding multiclass loss, leading to

$$\ell^*_{\mathrm{PAL}}(\mathbf{y}^*, \hat{\mathbf{y}}) = \sum_{j: y^*_j = 1} \ell_{\mathrm{MC}}(j, \hat{\mathbf{y}}) = \sum_{j \in [l]} y^*_j \ell_{\mathrm{MC}}(j, \hat{\mathbf{y}}). \tag{2}$$

Both approaches are consistent for precision at k. In order to make the reductions consistent for recall instead of precision, the label value needs to be replaced with a normalized label

$$\tilde{y}_j^* \coloneqq \frac{y_j^*}{\sum_{i=1}^{l} y_i^*} = \frac{y_j^*}{1 + \sum_{i\neq j}^{l} y_i^*}, \tag{3}$$

where the expression on the right has the advantage of being well defined even if there are no positives for the sample. This leads to the OVA-N and PAL-N reductions. By moving label-independent parts into functions f and g_j, the reductions get a common structure

$$\ell^*(\mathbf{y}^*, \hat{\mathbf{y}}) = f(\hat{\mathbf{y}}) + \sum_{j=1}^{l} z_j^* g_j(\hat{\mathbf{y}}), \tag{4}$$

where $z_j = \tilde{y}_j^*$ for the normalized reductions and $z_j^* = y_j^*$ otherwise. The functions f and g_j are the same for the normalized and regular reduction (see appendix).

3 Unbiased Estimates with Missing Labels

We are interested in noisy labels where the noise is such that labels can only go missing. This is described by the next two definitions, where the first gives a phenomenological characterization of the setting, whereas the second defines the mathematical model used to describe it. For this setting we then develop unbiased estimates for the preceding loss reductions, in the sense that for a given loss ℓ^* we are looking for a new loss function ℓ such that $\mathbb{E}\Big[\ell(\mathbf{Y}, \hat{\mathbf{Y}})\Big] = \mathbb{E}\Big[\ell^*(\mathbf{Y}^*, \hat{\mathbf{Y}})\Big]$.

Definition 1 (Propensity)**.** The missing-labels setting we described informally in the introduction leads to the following conditions on the l random variables

$$\mathbb{P}\big\{Y_j = 1 \mid Y_j^* = 1, \mathbf{Y^*}_{\neg j}, X\big\} =: p_j(X), \qquad \mathbb{P}\big\{Y_j = 1 \mid Y_j^* = 0, \mathbf{Y^*}_{\neg j}, X\big\} = 0 \tag{5}$$

The value $p_j(x) \in (0, 1]$ is called the *propensity* of the label j at point x.

Such propensity models have been used in extreme classification [32, 16, 43], learning-to-rank [20, 29, 41], and recommendation systems [34, 14, 15].

The following proposition guarantees that a fixed-propensity unbiased estimator can be used to construct a instance-dependent unbiased estimator

Proposition 1. *Let $f^*(X, Y^*)$ be some function such that for fixed propensity $\mathbf{p}$, an unbiased estimate is given by $f_{\mathbf{p}}$, i.e. $\mathbb{E}[f_{\mathbf{p}}(X, Y)] = \mathbb{E}[f^*(X, Y^*)]$. For instance-dependent propensity $\mathbf{p}(x)$, an unbiased estimator of f^* is given by $f_{\mathbf{p}(X)}$.*

Proof. Using the law of total expectation gives

$$\mathbb{E}[f^*(X, Y^*)] = \mathbb{E}[\mathbb{E}[f^*(X, Y^*) \mid X]] = \mathbb{E}\big[\mathbb{E}\big[f_{p(X)}(X, Y^*) \mid X\big]\big] = \mathbb{E}\big[f_{p(X)}(X, Y^*)\big]\,. \quad \square$$

Therefore, we will supress the dependence of the propensity on the data point in the rest of the paper.

The relation between $\mathbf{Y}^*$ and $\mathbf{Y}$ can be modeled by a set of independent *mask* variables $\mathbf{M}$:

Theorem 1 (Masking Model)**.** *Assuming $\mathbf{Y}^*$ and $\mathbf{Y}$ follow Definition 1, then then there exists a random variable $\mathbf{M} \in \{0, 1\}^l$ such that $\mathbf{Y} = \mathbf{M} \odot \mathbf{Y}^*$ almost surely and M_j is independent of $(\mathbf{Y}^*, X, \mathbf{M}_{\neg j})$ for all $j \in [l]$. It holds that $\mathbb{E}[M_j] = p_j$.*

This can be seen as a multilabel generalization of the similar statement given in Teisseyre et al. [37]. The independent variables $\mathbf{M}$ provide a convenient framework for proving the results that follow, because the independence allows to factorize expectations containing $\mathbf{M}$.

Proposition 2 (Unbiased Estimate for Decomposable Reductions)**.** *Assume the setting of Definition 1, with the additional condition that the predictions $\hat{\mathbf{Y}}$ are independent of the missing mask $\mathbf{M}$. Then the unbiased estimate for the loss* (4) *with $z = y$, denoted by $\ell = \mathfrak{P}(\ell^*)$, is given by*

$$\ell(\mathbf{y}, \hat{\mathbf{y}}) = f(\hat{\mathbf{y}}) + \sum_{j=1}^{l} \frac{y_j}{p_j} g_j(\hat{\mathbf{y}}). \tag{6}$$

The predictions have to be independent of the locations $\mathbf{M}$ where the labels go missing. This is fulfilled if the predictions $\hat{\mathbf{Y}} = h(X, \mathbf{W})$ are the output of a classifier h whose weights $\mathbf{W}$ are independent of $\mathbf{M}$.[1]

For the normalized reductions, it would suffice to find an unbiased estimate of $\tilde{Y}$ in order to apply the same argument as above. However, we are not aware of a derivation for such an estimate that is simpler than the fully generic case presented below.

Theorem 2 (Unbiased Estimate for Non-Decomposable Loss)**.** *For a generic multilabel loss function ℓ^*, the unbiased estimate $\ell = \mathfrak{P}(\ell^*)$ under the conditions of Proposition 2 is given by*

$$\ell(\mathbf{y}, \hat{\mathbf{y}}) = \sum_{\mathbf{y}' \preceq \mathbf{y}} \prod_{j : y_j = 1} \left(\frac{y'_j (2 - p_j) + p_j - 1}{p_j} \right) \ell^*(\mathbf{y}', \hat{\mathbf{y}}), \tag{7}$$

where $\mathbf{y}' \preceq \mathbf{y}$ means $\{0, 1\} \ni y'_j \leq y_j$.

This means that for an instance with k positive labels, we need 2^k evaluations of the original loss function in order to calculate the unbiased estimate. This is only feasible because, despite having a very large label space, typical extreme-classification datasets have only few positives per instance.

Unfortunately, the division by (products of) propensity values means that the unbiased estimates will have much larger variance than the original loss function would have on clean data. As an illustrative example, consider the binary case in the limit $p \ll 1$. We can show that in this case the variance grows with p^{-1} compared to the evaluation on clean data.

Proposition 3 (Increase in Variance)**.** *Setting $q^* := \mathbb{E}[Y^*]$ and $\ell = \mathfrak{P}(\ell^*)$, for small propensities $p \ll 1$, the variance increases with the inverse of the propensity, $\mathbb{V}[\ell(Y, \hat{y})] \approx \frac{1}{p(1-q^*)} \mathbb{V}[\ell^*(Y^*, \hat{y})]$.*

This means that in the binary case the variance increases linearly with inverse propensity. In the multilabel case, this is amplified further due to the product of propensities.

The result above raises the question whether there might be other unbiased estimators with reduced variance. For example, the conditional expectation $\mathbb{E}[\ell^*(Y^*, X) | Y]$ also gives an unbiased estimate with lower variance, but cannot be calculated without knowledge of the conditional probabilities $\mathbb{P}\{Y \mid X\}$. The following theorem states that $\ell = \mathfrak{P}(\ell^*)$ is unique if we want the loss function to work for all possible distributions of data. Thus we cannot reduce the variance.

Theorem 3 (Uniqueness)**.** *Let $p_j \in (0, 1]\ \forall j \in [l]$. For an arbitrary loss function ℓ^*, let ℓ and ℓ' be unbiased versions, in the sense that for all $X, \mathbf{Y}, \mathbf{Y}^*$ that fulfill the masking model Theorem 1 with propensity $\mathbf{p}$, it holds*

$$\mathbb{E}[\ell^*(\mathbf{Y}^*, X)] = \mathbb{E}[\ell(\mathbf{Y}, X)] = \mathbb{E}[\ell'(\mathbf{Y}, X)] \,. \tag{8}$$

Then, $\ell' = \ell$.

The unavoidable increase in variance indicates that there might be a bias-variance trade-off between using the unbiased loss that may overfit more strongly on the observed noise, and using the original loss function which gives wrong results even if $n \to \infty$. If one calculates a standard Rademacher bound for generalization (see appendix), this error bound increases with a factor $\frac{2-p}{p}$. [2]

In a classical learning setup, the generalization error would be described by the difference between the empirical risk and the true risk $\hat{\mathrm{R}}^*_{\ell^*}\left[\hat{h}\right] - \mathrm{R}^*_{\ell^*}\left[\hat{h}\right]$. However, in the case of missing labels, this can be decomposed in two ways

$$\begin{aligned} \mathrm{R}^*_{\ell*}[h] - \hat{\mathrm{R}}_{\ell}[h] &= \overbrace{\mathrm{R}^*_{\ell*}[h] - \mathrm{R}_{\ell}[h]}^{=0} + \mathrm{R}_{\ell}[h] - \hat{\mathrm{R}}_{\ell}[h] & (9) \\ &= \underbrace{\mathrm{R}^*_{\ell*}[h] - \hat{\mathrm{R}}^*_{\ell*}[h]}_{\text{finite sample}} + \underbrace{\hat{\mathrm{R}}^*_{\ell*}[h] - \hat{\mathrm{R}}_{\ell}[h]}_{\text{noise pattern}}, & (10) \end{aligned}$$

Whereas the first equation is just a restatement of the unbiasedness, the second contains some new insight: The generalization error can be decomposed into the difference between the true risk $\mathrm{R}^*_{\ell*}[h]$

[1] In this sense, we will use the notation $\ell(y, x)$ to evaluate a loss also on a data point.

[2] The bound in this paper corresponds to Natarajan et al. [28, Thm. 9], though that published result is wrong and missing the increase in the bound due to the increased range of the function.

and the empirical risk on clean training data $\hat{\mathrm{R}}^*_{\ell*}[h]$, and the difference between that and the estimated empirical risk on observed data $\hat{\mathrm{R}}_{\ell}[h]$. Because the classifier h depends (through $Y = \mathbf{M} \odot Y^*$) on the mask variables, ℓ does not give an unbiased estimate (on training data) and thus the second term is non-zero even in expectation. In fact, in the low-regularization regime this term may dominate the entire error, as we will demonstrate in section 5.

4 Convex Upper-Bounds

The unbiased estimate allows us to calculate the loss even on data with missing labels, but can we also use it for training? Ideally, the loss function should be lower-bounded, so the minimization is well defined, it should be convex so the minimum is unique. Further, the variance of the unbiased estimator should not be too large, so that a reasonable amount of training samples is sufficient.

If we assume ℓ_{BC} and ℓ_{MC} to be lower-bounded and convex, then only the PAL-reduction results in an unbiased estimate that is guaranteed to have the same properties, as it is a positive combination of ℓ_{MC}. Due to the uniqueness result, it is not possible to find an unbiased estimate that is always convex for the other reductions. Thus, in order to make them amenable for training, we propose to switch from unbiased estimates to convex upper-bounds. Below we present solutions for the OvA and normalized PAL-reduction. The normalized OVA-reduction remains an open problem.

Upper-Bound for OvA-Reduction The OvA-reduction is based on a binary loss, which often is a convex surrogate for the 0-1 loss. To get a convex loss in the missing-labels case, we thus switch the order of operations [32, 5]: Instead of taking an unbiased estimate of a convex surrogate, we form a convex surrogate of an unbiased estimate. Taking θ to be a thresholding function (e.g. $\theta(s) = \mathbb{1}\{s > 0\}$), the 0-1-loss can be written as

$$\ell^*_{0-1}(y, \hat{y}) = y\theta(\hat{y}) + (1-y)(1-\theta(\hat{y})) \tag{11}$$

with unbiased estimate

$$\ell_{0-1}(y, \hat{y}) = \left(\frac{2}{p_j} - 1\right) y\theta(\hat{y}) + (1-y)(1-\theta(\hat{y})) + y\left(\frac{p_j - 1}{p_j}\right). \tag{12}$$

As the last term does not depend on the predictions, it can be dropped for an optimization objective. If $\ell_{\mathrm{BC}}(1, \hat{y})$ is a convex upper-bound on $\theta(\hat{y})$ and $\ell_{\mathrm{BC}}(0, \hat{y})$ on $(1 - \theta(\hat{y}))$, so that overall ℓ_{BC} is a convex upper-bound on the 0-1 loss, then performing these substitutions gives a convex loss function for the OvA-reduction:

$$\tilde{\ell}_{\mathrm{OvA}}(\mathbf{y}, \hat{\mathbf{y}}) = \sum_{j=1}^{l} \left(\frac{2}{p_j} - 1\right) y_j \ell_{\mathrm{BC}}(1, \hat{y}_j) + (1-y_j)\ell_{\mathrm{BC}}(0, \hat{y}_j) \tag{13}$$

Upper-Bound for Normalized PAL-Reduction We have formulated the normalized multilabel reductions in terms of the variable $\tilde{Y}^*$. A naive attempt of correcting for the noisy labels by replacing Y^* with Y/p is not unbiased. However, the resulting estimator $\tilde{Y}$ turns out to be an upper bound. The two estimators are given by

$$\tilde{Y}^*_i = \frac{Y^*_i}{1 + \sum_{j \neq i} Y^*_j}, \qquad \tilde{Y}_i \coloneqq \frac{Y_i/p_i}{1 + \sum_{j \neq i} Y_j/p_j}. \tag{14}$$

Theorem 4 (Normalized Label Upper-Bound)**.** *Under the conditions of Theorem 2, replacing the true label with the unbiased estimate of the observed label as shown in Equation 14 results in an upper bound, whose error itself can be bounded by a data-dependent term*

$$\mathbb{E}\left[\tilde{Y}^*_i\right] + \sum_{j \neq i} \left(\frac{1-p_j}{p_j}\right) \mathbb{E}\left[\frac{Y_i}{p_i} \cdot \frac{Y_j}{p_j}\right] \geq \mathbb{E}\left[\tilde{Y}_i\right] \geq \mathbb{E}\left[\tilde{Y}^*_i\right]. \tag{15}$$

Proof. For convenience denote $S^*_i \coloneqq \sum_{j \neq i} Y^*_j$ and $S_i \coloneqq \sum_{j \neq i} Y_j/p_j$, and note that S_i is independent of M_i. By pulling out known factors and using the independence of M and $\mathbf{Y}^*$ we can show that

$$\mathbb{E}[S_i \mid \mathbf{Y}^*] = \sum_{j \neq i} \mathbb{E}\big[M_j Y^*_j / p_j \mid \mathbf{Y}^*\big] = \sum_{j \neq i} Y^*_j \, \mathbb{E}[M_j/p_j \mid \mathbf{Y}^*] = S^*_i. \tag{16}$$

Expanding terms and using independence of M_i, then applying the tower property and pulling out the measurable factor results in

$$\mathbb{E}\Big[\tilde{Y}_i\Big] = \mathbb{E}\left[\frac{M_i Y_i^*/p_i}{1+S_i}\right] = \mathbb{E}\left[\frac{M_i}{p_i}\right]\mathbb{E}\left[\frac{Y_i^*}{1+S_i}\right] = \mathbb{E}\left[\mathbb{E}\left[\frac{Y_i^*}{1+S_i} \;\middle|\; \mathbf{Y}^*\right]\right] = \mathbb{E}\left[Y_i^*\,\mathbb{E}\left[\frac{1}{1+S_i} \;\middle|\; \mathbf{Y}^*\right]\right].$$

The function $h : \mathbb{R}_{\geq 0} \longrightarrow \mathbb{R}$ given by $t \mapsto 1/(1+x)$ is convex, because its second derivative is $2(1+t)^{-3}$, which is larger than zero for non-negative t. Because $S_i \geq 0$ almost surely, we can apply Jensen's inequality to the inner expectation and use (16)

$$\mathbb{E}\Big[\tilde{Y}_i\Big] \geq \mathbb{E}\left[\frac{Y_i^*}{1+\mathbb{E}[S_i \mid \mathbf{Y}^*]}\right] = \mathbb{E}\left[\frac{Y_i^*}{1+S_i^*}\right] = \mathbb{E}\Big[\tilde{Y}_i^*\Big].$$

On the other hand, we can use the Taylor formula with intermediate point $\zeta \in [S_i, S_i^*]$ to expand

$$\frac{1}{1+S_i} = \frac{1}{1+S_i^*} - \frac{S_i - S_i^*}{(1+S_i^*)^2} + \frac{(S_i - S_i^*)^2}{(1+\zeta)^3}. \tag{17}$$

Using $\zeta \geq 0$ to bound the denominator, then multiplying with Y_i^* and taking the expectation gives

$$\mathbb{E}\left[\frac{Y_i^*}{1+S_i}\right] \leq \mathbb{E}\left[\frac{Y_i^*}{1+S_i^*}\right] + \mathbb{E}\big[Y_i^*(S_i - S_i^*)^2\big]. \tag{18}$$

The variance term can be calculated by substituting S_i and S_i^*, expanding the sum, and using the independence of M to show that the mixed terms are zero:

$$\begin{aligned}\mathbb{E}\big[Y_i^*(S_i - S_i^*)^2\big] &= \mathbb{E}\left[Y_i^*\left(\sum_{j\neq i} Y_j^*\left(\frac{M_j}{p_j} - 1\right)\right)^2\right] \\ &= \sum_{j\neq i}\mathbb{E}\left[Y_i^*(Y_j^*)^2\left(\frac{M_j}{p_j} - 1\right)^2\right] + \sum_{j\neq i}\sum_{k\notin\{i,j\}}\mathbb{E}\big[Y_i^* Y_j^* Y_k^*\big]\,\mathbb{E}\left[\frac{M_j}{p_j} - 1\right]\mathbb{E}\left[\frac{M_k}{p_k} - 1\right] \\ &= \sum_{j\neq i}\mathbb{E}\big[Y_i^* Y_j^*\big]\,\mathbb{E}\left[\frac{M_j}{p_j^2} - 2\frac{M_j}{p_j} + 1\right] = \sum_{j\neq i}\left(\frac{1-p_j}{p_j}\right)\mathbb{E}\left[\frac{Y_i}{p_i}\cdot\frac{Y_j}{p_j}\right]. \quad \square\end{aligned} \tag{19}$$

Note that the transformation of equation (3) was crucial for this calculation, because it makes the mask variables in the numerator and denominator independent.

Table 1: Error bound for XMC datasets

Dataset	Average	Worst Case
Eurlex-4K	0.02	0.51
AmazonCat-13K	0.0006	0.24

In practice, most entries of the co-occurrence matrix $\mathbb{E}[Y_i \cdot Y_j]$ will be extremely small, causing only a minute contribution to the error bound. This can be illustrated by calculating, on two real datasets, the upper-bound for the error of the proposed estimator, by approximating $\mathbb{E}[Y_i \cdot Y_j]$ with the label co-occurrence frequency. The propensities are estimated as in Jain et al. [16]. Looking at the mean value, and the worst case for any label (Table 1), We can see that the error on average is very small, indicating that the worst-case bound only applies to very few labels.

Corollary 1 (PAL Upper-Bound)**.** *Under the assumptions of Theorem 2, if the underlying multiclass loss ℓ_{MC} is a non-negative convex function, the expression*

$$\tilde{\ell}(\mathbf{y}, \hat{\mathbf{y}}) \coloneqq \sum_{j=1}^{l} \frac{y_i/p_i}{1+\sum_{j\neq i} y_j/p_j}\,\ell_{MC}(j, \hat{\mathbf{y}}) \tag{20}$$

gives a nonnegative, convex upper-bound on the true normalized PAL loss in expectation.

5 Experimental Results

In this section we present some empirical evidence that illustrates the influence of missing labels and the unbiased estimates and upper bounds on overfitting and bias-variance trade-off. Additional results and a more detailed description of the procedure can be found in the appendix.

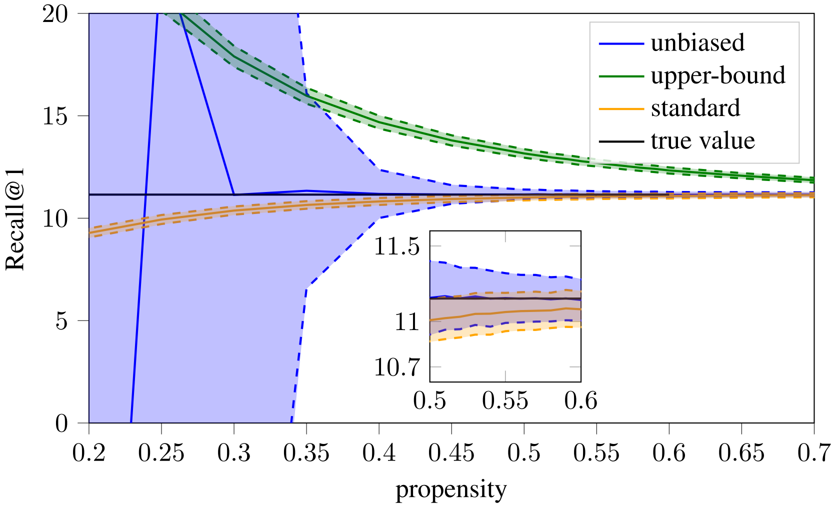

Figure 1: Unbiased estimate of per-example recall with artificial data as described in the main text. The shaded region corresponds to one standard deviation, estimated over 100 repetitions. The black line denotes the true recall.

Prediction Setting First, we want to demonstrate the variance problem in a simple prediction setting, where the classifier is fixed and we want to determine its performance. Consider a setting in which there are 100 different labels, which are independent and each has a probability of 10%. We randomly draw $10\,000$ ground-truth label vectors, and generate observed labels by removing according to a propensity p that is identical for all labels. The predictions are generated by randomly choosing a label from the ground-truth. We calculate the average per-example recall using the standard estimator, the unbiased estimator, and the upper bound, and plot the results in Figure 1.

As can be seen, for moderate propensities the unbiased estimator works well, but for propensities below 0.45 the $10\,000$ samples are not sufficient to get an accurate estimate. In this setting, the upper-bound results in a larger error than using the standard estimator.

Training Setting Ideally, we would benchmark our loss functions on a real XMC task. However, for those we neither know the exact propensities, nor can we validate that the unbiased estimates and upper bounds produce reasonable results, since the fully-labeled ground truth is unknown.

Instead of using fully artificial data, we chose to construct a dataset based on existing data: We took `AmazonCat-13k`[23] and consider only the 100 most common labels, which are the ones with the highest propensity according to Jain et al. [16]. We artificially remove labels according to inverse propensity, which increases linearly based on the ordering of label frequencies, such that the most common label has an inverse propensity of 2 and the 100th most common one of 20. This process partially preserves the strong imbalances that are typical of extreme classification datasets.

On this data, we train a linear classifier with L_2-regularization using different basis loss functions with **a)** the original (standard) loss on clean training data and **b)** noisy training data, as well as **c)** the unbiased version and **d)** the upper-bound version on noisy data. For each training run, we evaluate the loss on noisy and clean training and test data. For the evaluation on noisy data, the corresponding unbiased estimators are used.

In this linear-classifier experiment, the noise-pattern overfitting is much stronger than the overfitting due to finite sampling (10). Figure 2 shows this for the case of the BCE loss in OvA-reduction and CCE loss in normalized PAL reduction. For the classifier trained on clean data (blue), the weights are independent of the noise pattern and thus the dashed and dotted lines coincide in expectation. For the case of OvA reduction using the BCE loss, the training loss gets reduced much further using the unbiased loss function or the upper-bound loss function than using the standard loss. This decrease more than compensates the increase in generalization gap, and as such the minimal test loss is better with these two variants of the loss function. In contrast, in the non-decomposable case, even though

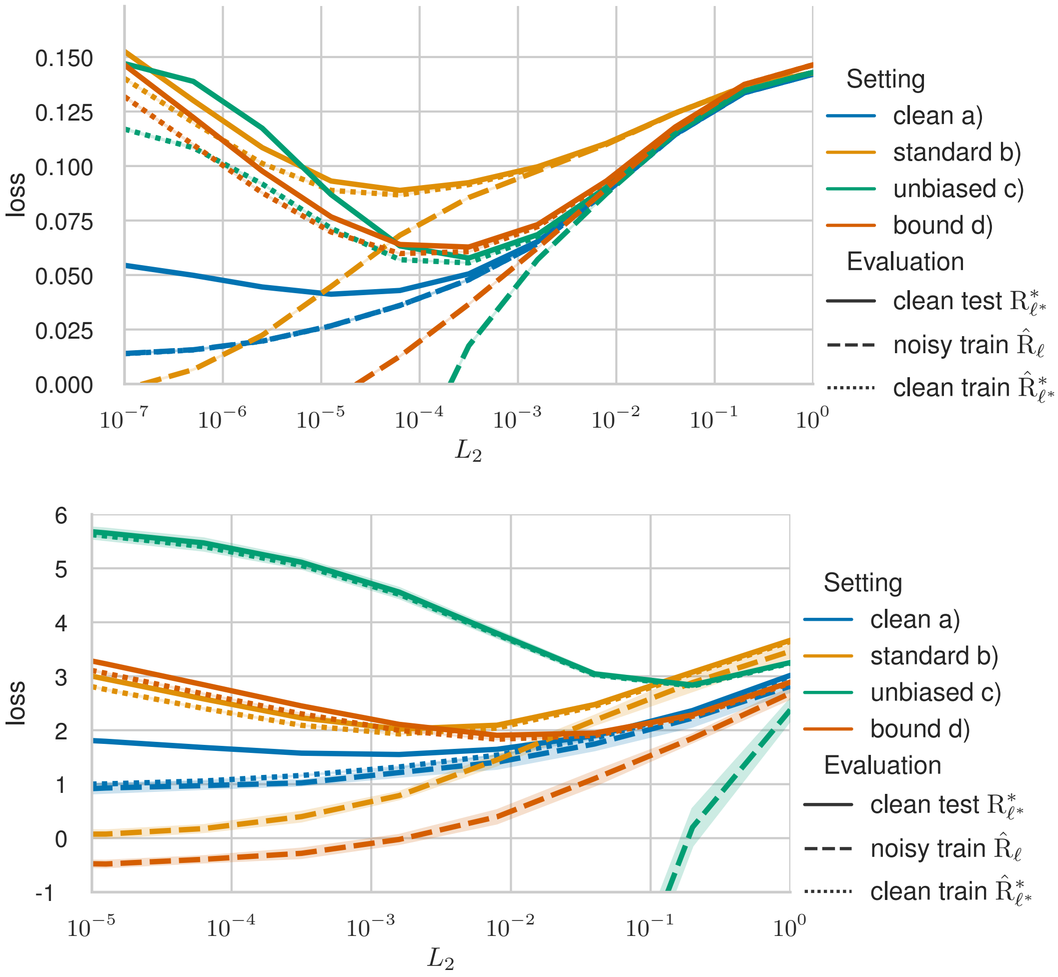

Figure 2: Binary cross-entropy (top) and normalized categorical cross-entropy (bottom) for different regularization strengths, evaluated on noisy training data, clean training data, and clean test data. The gaps between dashed and dotted lines correspond to overfitting to the noise pattern, the smaller gaps between dotted and solid lines show the generalization gaps due to the finite training sample. As the dashed lines are for noisy data, they are calculated using the unbiased estimate (6).

the observed training loss decreases drastically with the unbiased loss, the increase in overfitting makes the test loss worse than using the biased standard loss function.

In this case, using the upper-bound (20) can mitigate the effect, though there is still significant overfitting, as evidenced by the estimated training loss being less than zero. This is possible because even though the loss we use for training is a non-negative upper bound on the expected unbiased loss, the dashed curves show the value estimated for the loss using the unbiased estimator, which can be negative due to overfitting. For the OVA case, the upper bound (13) also reduces overfitting, but does not result in an overall better classifier on test data.

In terms of the bias-variance trade-off, the graphs show a clear trend: The optimal regularization for training on noisy data is larger than on clean data. It is also larger when using the unbiased or upper-bound loss as compared to standard loss. This is as expected from the variance analysis and generalization bound presented in the theory.

6 Related Work

Unbiased Estimates for Noisy Labels Learning with missing labels is a specific instance of learning with corrupted labels. For the case of binary labels, unbiased estimates of the loss function can be found in Natarajan et al. [28]. A more general approach is given in Van Rooyen and Williamson [39]. In their notation, f is a function and $\mathbb{P}$ the probability distribution over clean data, that is transformed by the invertible operator T into a *corrupted* probability distribution. Let R be the inverse of T, and R^* its adjoint, then $\langle \mathbb{P}, f\rangle = \langle \mathsf{R} \circ \mathsf{T}(\mathbb{P}), f\rangle = \langle \mathsf{T}(\mathbb{P}), \mathsf{R}^*(f)\rangle$. This equation forms the basis for their "Theorem 5 (Corruption Corrected Loss)", which states that a *corruption corrected* function l_{R} is given $\forall a \in \mathcal{A}$ by $l_{\mathrm{R}}(\cdot, a) = \mathsf{R}^*(l(\cdot, a))$, where $\mathcal{A}$ denotes the set of possible actions that will be evaluated by the loss functions. For a finite label space with n possible, the operator R^* can be represented with an $n \times n$ matrix. For the multilabel case here, applying this naively would require 2^l evaluations of the original loss function. In contrast, the direct approach presented in section 3 is much more efficient.

Alternatives In some settings with noisy labels, it is possible to use a learning algorithm that is inherently noise tolerant [12, 40]. Certain performance objectives such as the balanced error or the AUC are noise robust even under the more general setting of mutually contaminated distributions as shown in Menon et al. [24]. A data re-calibration approach tries to identify from the training data which samples are corrupted, e.g. by looking at samples for which the network is very unsure, and adapt the training process correspondingly [13, 46, 19] It is also possible to first train a scorer on the noisy data naively, from which a classifier adapted to a given rate of missing labels can be constructed by choosing an appropriate threshold [24]. Similarly, the inference procedure of PLTs can be adapted to take into account a propensity model [43].

Related Learning Settings Learning with missing labels is highly related to learning from positive and unlabeled (PU) data [11]. An unbiased loss function for this setting is given in Du Plessis et al. [10]. The appearing difficulties, that non-negativity and convexity need not be preserved, are the same as in our setting [22]. A slightly different setting with missing labels is given by semi-supervised learning, where it is know for which labels are missing [45].

7 Summary and Discussion

This paper provides unbiased estimates for four cases of multilabel reductions given in Menon et al. [25]. Except for the PAL reduction, these estimators can be non-convex and even negatively unbounded. The unbiased estimates come with an increase in variance. This is unavoidable if unbiasedness is required, as the estimators can be shown to be unique. If sufficient training data is available, then the unbiased loss functions can be used, but for the normalized reductions we found that even 1.2 million instances in AmazonCat are not enough. Much fewer data points are needed in order to estimate the overall loss of a classifier. This is because for training, an accurate estimate for $\mathbb{E}[\ell(Y^*, h(X) \mid X]$ needs to be formed, whereas for evaluation this is averaged over the entire dataset, $\mathbb{E}[\ell(Y^*, h(X)]$. This indicates that the unbiased estimates can be useful for hyperparameter tuning and model selection.

For training, however, another approach is needed. A method that fixes the negative unboundedness and non-convexity and also reduces the variance is to switch to a convex upper-bound. We have shown that this can stabilize the training and improve the results.

Furthermore, the data in section 5 suggest training with missing labels requires more regularization, irrespective of whether training uses standard-, unbiased-, or convex upper-bound losses. Our findings agree with Arpit et al. [2] who found that typical regularizers prevent a deep network from memorizing *noisy* examples, while not hindering the learning of patterns from *clean* instances.

All in all, our results show that a) unbiasedness can be achieved for generic multilabel losses, and in particular the losses resulting from multilabel reduction, but also that b) these losses might not be suitable for optimization. We have presented one method that trades away unbiasedness for the ability to handle training with lower amounts of data. An exciting future research prospect would be to investigate families of loss functions that can continuously trade off bias and variance, and thus allow for optimal training with different amounts of available data.

References

[1] R. Agrawal, A. Gupta, Y. Prabhu, and M. Varma. Multi-label learning with millions of labels: Recommending advertiser bid phrases for web pages. In *Proceedings of the 22nd International Conference on World Wide Web*, WWW '13, page 13–24, New York, NY, USA, 2013. Association for Computing Machinery.

[2] D. Arpit, S. Jastrzębski, N. Ballas, D. Krueger, E. Bengio, M. S. Kanwal, T. Maharaj, A. Fischer, A. Courville, Y. Bengio, and S. Lacoste-Julien. A Closer Look at Memorization in Deep Networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 233–242. PMLR, 06–11 Aug 2017.

[3] R. Babbar and B. Schölkopf. Dismec: Distributed sparse machines for extreme multi-label classification. In *WSDM*, pages 721–729, 2017.

[4] K. Bhatia, K. Dahiya, H. Jain, Y. Prabhu, and M. Varma. The extreme classification repository: Multi-label datasets and code. `http://manikvarma.org/downloads/XC/XMLRepository.html`, 2016.

[5] Y.-T. Chou, G. Niu, H.-T. Lin, and M. Sugiyama. Unbiased Risk Estimators Can Mislead: A Case Study of Learning with Complementary Labels. In *International Conference on Machine Learning*, pages 1929–1938. PMLR, Nov. 2020.

[6] K. Dahiya, A. Agarwal, D. Saini, K. Gururaj, J. Jiao, A. Singh, S. Agarwal, P. Kar, and M. Varma. Siamesexml: Siamese networks meet extreme classifiers with 100m labels. In *Proceedings of the International Conference on Machine Learning*, July 2021.

[7] O. Dekel and O. Shamir. Multiclass-multilabel classification with more classes than examples. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 137–144, 2010.

[8] K. Dembczynski, W. Cheng, and E. Hüllermeier. Bayes optimal multilabel classification via probabilistic classifier chains. In *ICML*, 2010.

[9] J. Deng, A. C. Berg, K. Li, and L. Fei-Fei. What does classifying more than 10,000 image categories tell us? In *ECCV*, 2010.

[10] M. Du Plessis, G. Niu, and M. Sugiyama. Convex formulation for learning from positive and unlabeled data. In *International conference on machine learning*, pages 1386–1394, 2015.

[11] C. Elkan and K. Noto. Learning classifiers from only positive and unlabeled data. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '08, pages 213–220, New York, NY, USA, Aug. 2008. Association for Computing Machinery.

[12] A. Ghosh, N. Manwani, and P. S. Sastry. Making risk minimization tolerant to label noise. *Neurocomputing*, 160:93–107, 2015. Publisher: Elsevier.

[13] B. Han, Q. Yao, X. Yu, G. Niu, M. Xu, W. Hu, I. Tsang, and M. Sugiyama. Co-teaching: Robust training of deep neural networks with extremely noisy labels. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[14] J. Huang, H. Oosterhuis, M. De Rijke, and H. Van Hoof. Keeping dataset biases out of the simulation: A debiased simulator for reinforcement learning based recommender systems. In *Fourteenth ACM conference on recommender systems*, pages 190–199, 2020.

[15] J. Huang, H. Oosterhuis, and M. de Rijke. It is different when items are older: Debiasing recommendations when selection bias and user preferences are dynamic. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, pages 381–389, 2022.

[16] H. Jain, Y. Prabhu, and M. Varma. Extreme multi-label loss functions for recommendation, tagging, ranking and other missing label applications. In *KDD*, August 2016.

[17] H. Jain, V. Balasubramanian, B. Chunduri, and M. Varma. Slice: Scalable linear extreme classifiers trained on 100 million labels for related searches. In *WSDM*, pages 528–536, 2019.

[18] Y. Jernite, A. Choromanska, and D. Sontag. Simultaneous learning of trees and representations for extreme classification and density estimation. In D. Precup and Y. W. Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1665–1674. PMLR, 06–11 Aug 2017.

[19] L. Jiang, Z. Zhou, T. Leung, L.-J. Li, and L. Fei-Fei. MentorNet: Learning data-driven curriculum for very deep neural networks on corrupted labels. In J. Dy and A. Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2304–2313. PMLR, 10–15 Jul 2018.

[20] T. Joachims, A. Swaminathan, and T. Schnabel. Unbiased learning-to-rank with biased feedback. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*, pages 781–789, 2017.

[21] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization, 2017.

[22] R. Kiryo, G. Niu, M. C. du Plessis, and M. Sugiyama. Positive-Unlabeled Learning with Non-Negative Risk Estimator. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 1675–1685. Curran Associates, Inc., 2017.

[23] J. McAuley and J. Leskovec. Hidden factors and hidden topics: understanding rating dimensions with review text. In *Proceedings of the 7th ACM conference on Recommender systems*, pages 165–172, 2013.

[24] A. Menon, B. Van Rooyen, C. S. Ong, and B. Williamson. Learning from corrupted binary labels via class-probability estimation. In *International Conference on Machine Learning*, pages 125–134, 2015.

[25] A. K. Menon, A. S. Rawat, S. Reddi, and S. Kumar. Multilabel reductions: what is my loss optimising? *Advances in Neural Information Processing Systems*, 32, 2019.

[26] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[27] M. Mohri, A. Rostamizadeh, and A. Talwalkar. *Foundations of machine learning*. MIT press, 2018.

[28] N. Natarajan, I. S. Dhillon, P. Ravikumar, and A. Tewari. Cost-sensitive learning with noisy labels. *The Journal of Machine Learning Research*, 18(1):5666–5698, 2017. Publisher: JMLR. org.

[29] H. Oosterhuis and M. de Rijke. Policy-aware unbiased learning to rank for top-k rankings. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 489–498, 2020.

[30] I. Partalas, A. Kosmopoulos, N. Baskiotis, T. Artieres, G. Paliouras, E. Gaussier, I. Androutsopoulos, M.-R. Amini, and P. Galinari. Lshtc: A benchmark for large-scale text classification. *arXiv preprint arXiv:1503.08581*, 2015.

[31] Y. Prabhu and M. Varma. Fastxml: A fast, accurate and stable tree-classifier for extreme multi-label learning. In *KDD*, pages 263–272. ACM, 2014.

[32] M. Qaraei, E. Schultheis, P. Gupta, and R. Babbar. Convex Surrogates for Unbiased Loss Functions in Extreme Classification With Missing Labels. In *Proceedings of the Web Conference 2021*, pages 3711–3720, Ljubljana Slovenia, Apr. 2021. ACM.

[33] S. J. Reddi, S. Kale, F. Yu, D. Holtmann-Rice, J. Chen, and S. Kumar. Stochastic negative mining for learning with large output spaces. In K. Chaudhuri and M. Sugiyama, editors, *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 1940–1949. PMLR, 16–18 Apr 2019.

[34] N. Sachdeva, C.-J. Wu, and J. McAuley. On sampling collaborative filtering datasets. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, WSDM '22, page 842–850, New York, NY, USA, 2022. Association for Computing Machinery.

[35] E. Schultheis, M. Wydmuch, R. Babbar, and K. Dembczyński. On missing labels, long-tails and propensities in extreme multi-label classification. *arXiv preprint arXiv:2207.13186*, 2022.

[36] S. Shalev-Shwartz and S. Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

[37] P. Teisseyre, J. Mielniczuk, and M. Łazęcka. Different strategies of fitting logistic regression for positive and unlabelled data. In V. V. Krzhizhanovskaya, G. Závodszky, M. H. Lees, J. J. Dongarra, P. M. A. Sloot, S. Brissos, and J. Teixeira, editors, *Computational Science – ICCS 2020*, pages 3–17, Cham, 2020. Springer International Publishing. ISBN 978-3-030-50423-6.

[38] G. Tsoumakas and I. M. Katakis. Multi-label classification: An overview. *Int. J. Data Warehous. Min.*, 3:1–13, 2007.

[39] B. Van Rooyen and R. C. Williamson. A theory of learning with corrupted labels. *The Journal of Machine Learning Research*, 18(1):8501–8550, 2017. ISSN 1532-4435.

[40] B. Van Rooyen, A. Menon, and R. C. Williamson. Learning with symmetric label noise: The importance of being unhinged. In *Advances in Neural Information Processing Systems*, pages 10–18, 2015.

[41] X. Wu, H. Chen, J. Zhao, L. He, D. Yin, and Y. Chang. Unbiased learning to rank in feeds recommendation. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, pages 490–498, 2021.

[42] M. Wydmuch, K. Jasinska, M. Kuznetsov, R. Busa-Fekete, and K. Dembczynski. A no-regret generalization of hierarchical softmax to extreme multi-label classification. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[43] M. Wydmuch, K. Jasinska-Kobus, R. Babbar, and K. Dembczynski. *Propensity-Scored Probabilistic Label Trees*, page 2252–2256. Association for Computing Machinery, New York, NY, USA, 2021.

[44] R. You, Z. Zhang, Z. Wang, S. Dai, H. Mamitsuka, and S. Zhu. Attentionxml: Label tree-based attention-aware deep model for high-performance extreme multi-label text classification. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[45] H.-F. Yu, P. Jain, P. Kar, and I. Dhillon. Large-scale multi-label learning with missing labels. In *International conference on machine learning*, pages 593–601, 2014.

[46] S. Zheng, P. Wu, A. Goswami, M. Goswami, D. Metaxas, and C. Chen. Error-bounded correction of noisy labels. In *International Conference on Machine Learning*, pages 11447–11457. PMLR, 2020.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [No]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...

   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes] Some proofs are in the appendix

3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [No]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [No]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

A Proofs

In this section we present the theorems of the main paper in a slightly more rigorous way and with more detailed proofs. Throughout this section, we assume an abstract probability space $(\Omega, \mathcal{F}, \mathbb{P})$ on which all random variables are defined.

A.1 Masking Model

Theorem 1 (Masking Model)**.** *Let $l \in \mathbb{N}$, $\mathbf{p} \in (0,1]^l$. We are given three random vectors $\mathbf{Y}^* : \Omega \longrightarrow \{0,1\}^l$, $\mathbf{Y} : \Omega \longrightarrow \{0,1\}^l$ and $X : \Omega \longrightarrow \mathcal{X}$ which fulfill*

$$\forall j \in [l] :\ \mathbb{P}\big\{Y_j = 1 \mid Y_j^* = 1, \mathbf{Y}^*{}_{\neg j}, X\big\} = p_j \tag{21}$$

$$\forall j \in [l] :\ \mathbb{P}\big\{Y_j = 1 \mid Y_j^* = 0, \mathbf{Y}^*{}_{\neg j}, X\big\} = 0, \tag{22}$$

where $\mathbf{Y}^*{}_{\neg j} \coloneqq \{Y_k^* : [l] \ni k \neq j\}$.

Then there exists $\mathbf{M} : \Omega \longrightarrow \{0,1\}^l$ such that $\mathbf{Y} = \mathbf{M} \odot \mathbf{Y}^*$ almost surely and M_j is independent of $(\mathbf{Y}^*, X, \mathbf{M}_{\neg j})$ for all $j \in [l]$. It holds that $\mathbb{E}[M_j] = p_j$.

Proof. Choose $\mathbf{M}'$ as Bernoulli variables independent from X and Y^* with $\mathbb{P}\big\{M'_j = 1\big\} = p_j$. Define $\mathbf{M}$ through

$$\forall j \in [l] :\ M_j \coloneqq (1 - Y_j^*) M'_j + Y_j. \tag{23}$$

Equality: Calculating the difference between the observed labels and the masked labels gives

$$Y_j - Y_j^* M_j = Y_j - Y_j^*(1 - Y_j^*) M'_j - Y_j^* Y_j = Y_j - Y_j^* Y_j = Y_j(1 - Y_j^*). \tag{24}$$

We want to show this is zero with probability one:

$$\mathbb{P}\big\{Y_j = Y_j^* M_j\big\} = \mathbb{P}\big\{Y_j = 0 \vee Y_j^* = 1\big\} = 1 - \mathbb{P}\big\{Y_j = 1 \wedge Y_j^* = 0\big\} = 1 - \mathbb{P}\big\{Y_j = 1 \mid Y_j^* = 0\big\} = 1. \tag{25}$$

Independence: As M_j can take only two states, to show independence it is enough to look at the conditional probabilities for one of the cases. Because we have already proven that $M_j \in \{0,1\}$ a.s., and therefore only one of the summands can be one at the same time.

$$\begin{aligned} \mathbb{P}\big\{M_j = 1 \mid M'_{\neg j}, X, \mathbf{Y}^*\big\} &= \mathbb{P}\big\{(1 - Y_j^*) M'_j = 1 \vee Y_j = 1 \mid M'_{\neg j}, X, \mathbf{Y}^*\big\} \\ &= \mathbb{P}\big\{Y_j = 1 \mid M'_{\neg j}, X, \mathbf{Y}^*\big\} + \mathbb{P}\big\{(1 - Y_j^*) M'_j = 1 \mid M'_{\neg j}, X, \mathbf{Y}^*\big\} \end{aligned} \tag{26}$$

Using the tower property and pulling out known factors gives

$$\mathbb{P}\big\{Y_j = 1 \mid M'_{\neg j}, X, \mathbf{Y}^*\big\} = \mathbb{E}\big[\mathbb{E}\big[\mathbb{1}\{Y_j = 1\} \mid Y_j^*, X\big] \mid \mathbf{M}'_{\neg j}, X, \mathbf{Y}^*\big] \tag{27}$$

$$= \mathbb{E}\big[\mathbb{1}\big\{Y_j^* = 1\big\}\, p_j \mid \mathbf{M}'_{\neg j}, X, \mathbf{Y}^*\big] = p_j \mathbb{1}\big\{Y_j^* = 1\big\} = p_j Y_j^*. \tag{28}$$

Similarly, we can pull out the independent factor M'_j and known factor $1 - Y_j^*)$

$$\mathbb{P}\big\{(1 - Y_j^*) M'_j = 1 \mid M'_{\neg j}, X, \mathbf{Y}^*\big\} = \mathbb{E}\big[(1 - Y_j^*) M'_j \mid \mathbf{M}'_{\neg j}, X, \mathbf{Y}^*\big] \tag{29}$$

$$= \mathbb{E}\big[M'_j = 1\big]\, \mathbb{E}\big[(1 - Y_j^*) \mid \mathbf{M}'_{\neg j}, X, \mathbf{Y}^*\big] \tag{30}$$

$$= p_j(1 - Y_j^*). \tag{31}$$

Therefore, we get

$$\mathbb{E}[M_j] = \mathbb{P}\big\{M_j = 1 \mid \mathbf{M}'_{\neg j}, X, \mathbf{Y}^*\big\} = p_j, \tag{32}$$

thus proving the independence of M_j from $(\mathbf{M}'_{\neg j}, X, \mathbf{Y}^*)$. Because $\mathbf{M}_{\neg_j}$ is a measurable function of $(\mathbf{M}'_{\neg j}, X, \mathbf{Y}^*)$, this also proves the independence from $(\mathbf{M}_{\neg j}, X, \mathbf{Y}^*)$. □

A.2 Unbiasedness Proofs

Proposition 2 (Unbiased Estimate for Decomposable Reductions)**.** *Let $l \in \mathbb{N}$, $f : \mathbb{R}^l \longrightarrow \mathbb{R}$, $g : \mathbb{R}^l \longrightarrow \mathbb{R}^l$, $\mathbf{p} \in \mathbb{R}^l$, and define*

$$\ell^*(\mathbf{y}, \hat{\mathbf{y}}) \coloneqq f(\hat{\mathbf{y}}) + \sum_{j=1}^{l} y_j g_j(\hat{\mathbf{y}}) \tag{33}$$

$$\ell(\mathbf{y}, \hat{\mathbf{y}}) \coloneqq f(\hat{\mathbf{y}}) + \sum_{j=1}^{l} \frac{y_j}{p_j} g_j(\hat{\mathbf{y}}). \tag{34}$$

Then, for any $\mathbf{Y}$, $\mathbf{Y}^*$ such that

$$\forall j \in [l] :\ \mathbb{P}\Big\{ Y_j = 1 \mid Y_j^* = 1, \mathbf{Y^*}_{\neg j}, \hat{\mathbf{Y}} \Big\} = p_j \tag{35}$$

$$\forall j \in [l] :\ \mathbb{P}\Big\{ Y_j = 1 \mid Y_j^* = 0, \mathbf{Y^*}_{\neg j}, \hat{\mathbf{Y}} \Big\} = 0, \tag{36}$$

the function ℓ gives an unbiased estimate for the true expected loss ℓ^*:

$$\mathbb{E}\Big[\ell^*(\mathbf{Y}^*, \hat{\mathbf{Y}})\Big] = \mathbb{E}\Big[\ell(\mathbf{Y}, \hat{\mathbf{Y}})\Big]. \tag{37}$$

Proof. By Theorem 1 we can write $\mathbf{Y} = \mathbf{M} \odot \mathbf{Y}^*$. Thus

$$\mathbb{E}\Big[\ell(\mathbf{Y}, \hat{\mathbf{Y}})\Big] = \mathbb{E}\Big[f(\hat{\mathbf{Y}})\Big] + \sum_{j=1}^{l} \mathbb{E}\Big[M_j Y_j^* g_j(\hat{\mathbf{Y}})\Big] / p_j. \tag{38}$$

As M_j is an independent (from Y_j^* and $\hat{\mathbf{Y}}$) factor it can be pulled out

$$= \mathbb{E}\Big[f(\hat{\mathbf{Y}})\Big] + \sum_{j=1}^{l} \frac{\mathbb{E}[M_j]}{p_j} \mathbb{E}\Big[Y_j^* g_j(\hat{\mathbf{Y}})\Big] \tag{39}$$

$$= \mathbb{E}\Big[f(\hat{\mathbf{Y}})\Big] + \sum_{j=1}^{l} \mathbb{E}\Big[Y_j^* g_j(\hat{\mathbf{Y}})\Big] = \mathbb{E}\Big[\ell^*(\mathbf{Y}^*, \hat{\mathbf{Y}})\Big]. \tag{40}$$

□

Theorem 2 (Unbiased Estimate for Non-Decomposable Loss)**.** *In the same setting as 2, consider the generic function $\ell^* : \{0,1\}^l \times \mathbb{R}^l \longrightarrow \mathbb{R}$ and define*

$$\ell(\mathbf{y}, \hat{\mathbf{y}}) \coloneqq \sum_{\mathbf{y}' \preceq \mathbf{y}} \prod_{j : y_j = 1} \left(\frac{y'_j(2 - p_j) + p_j - 1}{p_j} \right) \ell^*(\mathbf{y}', \hat{\mathbf{y}}),$$

where $\mathbf{y}' \preceq \mathbf{y}$ means $\forall j \in [l] :\ \{0,1\} \ni y'_j \leq y_j$. Then

$$\mathbb{E}\Big[\ell^*(\mathbf{Y}^*, \hat{\mathbf{Y}})\Big] = \mathbb{E}\Big[\ell(\mathbf{Y}, \hat{\mathbf{Y}})\Big]. \tag{41}$$

Proof. Given that the label space is discrete, the function ℓ^* can be rearranged to be linear in the labels

$$\ell^*(\mathbf{y}^*, \hat{\mathbf{y}}) = \sum_{\mathbf{y}' \in \{0,1\}^l} \mathbb{1}\{\mathbf{y}^* = \mathbf{y}'\}\, \ell^*(\mathbf{y}', \hat{\mathbf{y}}), \tag{42}$$

with the indicator function given by the product

$$\mathbb{1}\{\mathbf{y}^* = \mathbf{y}'\} = \prod_{j=1}^{l} \mathbb{1}\{y_j^* = y'_j\} = \prod_{j=1}^{l} \left(y_j^* y'_j + (1 - y_j^*)\left(1 - y'_j\right) \right). \tag{43}$$

For a given $\mathbf{y}'$ define the $\mathbf{Y}^*$ measurable random variable $G_k(\mathbf{y}')$ and the $\mathbf{Y}^*, \mathbf{M}_{\neg k}$ measurable variable $H_k(\mathbf{y}')$ as

$$G_k \coloneqq \prod_{j=1}^{k} \left(Y_j^* y_j' + \left(1 - Y_j^*\right)\left(1 - y_j'\right)\right), \quad H_k \coloneqq \prod_{i=k+1}^{l} \left(\frac{Y_i^* M_i}{p_i} y_i' + \left(1 - \frac{Y_i^* M_i}{p_i}\right)(1 - y_i')\right). \tag{44}$$

Now we can perform the following induction. By the tower property

$$\mathbb{E}\Big[G_k H_k \mid \hat{\mathbf{Y}}\Big] = \mathbb{E}\Big[\mathbb{E}\Big[G_k \left(\frac{Y_{k+1}^* M_{k+1}}{p_{k+1}} y_{k+1}' + \left(1 - \frac{Y_{k+1}^* M_{k+1}}{p_{k+1}}\right)(1 - y_{k+1}')\right) H_{k+1} \mid \mathbf{Y}^*, \mathbf{M}_{\neg k+1}, \hat{\mathbf{Y}}\Big] \mid \hat{\mathbf{Y}}\Big].$$

Pulling out measurable factors gives

$$\begin{aligned}
&= \mathbb{E}\Big[G_k H_{k+1}\, \mathbb{E}\Big[\left(\frac{Y_{k+1}^* M_{k+1}}{p_{k+1}} y_{k+1}' + \left(1 - \frac{Y_{k+1}^* M_{k+1}}{p_{k+1}}\right)(1 - y_{k+1}')\right) \mid \mathbf{Y}^*, \mathbf{M}_{\neg k+1}, \hat{\mathbf{Y}}\Big] \mid \hat{\mathbf{Y}}\Big] \\
&= \mathbb{E}\left[G_k H_{k+1} \left(\frac{Y_{k+1}^*\, \mathbb{E}\Big[M_{k+1} \mid \mathbf{Y}^*, \mathbf{M}_{\neg k+1}, \hat{\mathbf{Y}}\Big]}{p_{k+1}} y_{k+1}' + \right.\right. \\
&\qquad \left.\left. \left(1 - \frac{Y_{k+1}^*\, \mathbb{E}\Big[M_{k+1} \mid \mathbf{Y}^*, \mathbf{M}_{\neg k+1}, \hat{\mathbf{Y}}\Big]}{p_{k+1}}\right)(1 - y_{k+1}')\right) \mid \hat{\mathbf{Y}}\right].
\end{aligned}$$

Due to the independence of M_{k+1} and $\mathbb{E}[M_{k+1}] = p_{k+1}$ this simplifies to

$$\begin{aligned}
&= \mathbb{E}\Big[G_k H_{k+1} \left(\frac{Y_{k+1}^*\, \mathbb{E}[M_{k+1}]}{p_{k+1}} y_{k+1}' + \left(1 - \frac{Y_{k+1}^*\, \mathbb{E}[M_{k+1}]}{p_{k+1}}\right)(1 - y_{k+1}')\right) \mid \hat{\mathbf{Y}}\Big] \\
&= \mathbb{E}\Big[G_k H_{k+1} \left(Y_{k+1}^* y_{k+1}' + \left(1 - Y_{k+1}^*\right)\left(1 - y_{k+1}'\right)\right) \mid \hat{\mathbf{Y}}\Big] = \mathbb{E}\Big[G_{k+1} H_{k+1} \mid \hat{\mathbf{Y}}\Big].
\end{aligned}$$

From this follows

$$\begin{aligned}
\mathbb{E}\left[\prod_{i=1}^{l} \left(\frac{Y_i}{p_i} y_i' + \left(1 - \frac{Y_i}{p_i}\right)(1 - y_i')\right) \mid \hat{\mathbf{Y}}\right] &= \mathbb{E}\Big[H_0 \mid \hat{\mathbf{Y}}\Big] = \mathbb{E}\Big[G_0 H_0 \mid \hat{\mathbf{Y}}\Big] \\
&= \mathbb{E}\Big[G_l H_l \mid \hat{\mathbf{Y}}\Big] = \mathbb{E}\Big[G_l \mid \hat{\mathbf{Y}}\Big] = \mathbb{E}\left[\prod_{j=1}^{l} \left(Y_j^* y_j' + \left(1 - Y_j^*\right)\left(1 - y_j'\right)\right) \mid \hat{\mathbf{Y}}\right].
\end{aligned} \tag{45}$$

Note that the term inside the expectation on the left side $G_l(\mathbf{y}')$ is always zero if for any $j \in [l]$ it holds that $Y_i = 0$ but $y_i' = 1$. Therefore it suffices to sum over all subsets of the observed labels, $\mathbf{y}' \preceq \mathbf{y}$:

$$\begin{aligned}
\mathbb{E}\Big[\ell^*(\mathbf{Y}^*, \hat{\mathbf{Y}}) \mid \hat{\mathbf{Y}}\Big] &= \sum_{\mathbf{y}' \in \{0,1\}^l} \mathbb{E}\Big[\mathbb{1}\{\mathbf{Y}^* = \mathbf{y}'\}\, \ell^*(\mathbf{y}', \hat{\mathbf{Y}}) \mid \hat{\mathbf{Y}}\Big] = \sum_{\mathbf{y}' \in \{0,1\}^l} \mathbb{E}\Big[G_l(\mathbf{y}') \mid \hat{\mathbf{Y}}\Big]\, \ell^*(\mathbf{y}', \hat{\mathbf{Y}}) \\
= \sum_{\mathbf{y}' \preceq \mathbf{Y}} \mathbb{E}\Big[G_l(\mathbf{y}') \mid \hat{\mathbf{Y}}\Big]\, \ell^*(\mathbf{y}', \hat{\mathbf{Y}}) &= \sum_{\mathbf{y}' \preceq \mathbf{Y}} \mathbb{E}\left[\prod_{i=1}^{l} \left(\frac{Y_i}{p_i} y_i' + \left(1 - \frac{Y_i}{p_i}\right)(1 - y_i')\right) \mid \hat{\mathbf{Y}}\right] \ell^*(\mathbf{y}', \hat{\mathbf{Y}}) \\
&= \sum_{\mathbf{y}' \preceq \mathbf{Y}} \mathbb{E}\left[\prod_{i : Y_i = 1}^{l} \left(\frac{1}{p_i} y_i' + \left(1 - \frac{1}{p_i}\right)(1 - y_i')\right) \mid \hat{\mathbf{Y}}\right] \ell^*(\mathbf{y}', \hat{\mathbf{Y}}) \\
&= \sum_{\mathbf{y}' \preceq \mathbf{Y}} \mathbb{E}\left[\prod_{i : Y_i = 1}^{l} \frac{y_i'(2 - p_i) + p_i - 1}{p_i} \mid \hat{\mathbf{Y}}\right] \ell^*(\mathbf{y}', \hat{\mathbf{Y}}) \\
&= \mathbb{E}\left[\sum_{\mathbf{y}' \preceq \mathbf{Y}} \left(\prod_{i : Y_i = 1}^{l} \frac{y_i'(2 - p_i) + p_i - 1}{p_i}\right) \ell^*(\mathbf{y}', \hat{\mathbf{Y}}) \mid \hat{\mathbf{Y}}\right],
\end{aligned} \tag{46}$$

where we have used that if $Y_i = 0$ and $y_i' = 0$, the corresponding factor is 1 so it can be left out of the product. Taking the expectation on both sides of the equation proves the claim. □

A.3 Properties of Unbiased Estimators

Remark 1 (Increase in Variance). For a single binary label, $l = 1$ such that $q^* := \mathbb{E}[Y^*]$ and $\mathbb{E}[Y = 1 \mid Y^* = 1] = p$, let $\ell^* : \{0,1\} \times \mathbb{R} \longrightarrow \mathbb{R}$ be a loss function. Denote $\ell = \mathfrak{P}(\ell^*)$ for its unbiased estimate according to Theorem 2. For small propensities $p \ll 1$, the variance of the true loss and the estimate are related through

$$\mathbb{V}[\ell(Y, \hat{y})] \approx \frac{1}{p(1-q^*)} \mathbb{V}[\ell^*(Y^*, \hat{y})]\,. \tag{47}$$

Proof. For convenience denote $\ell^*_+(\hat{y}) = \ell^*(1, \hat{y})$ and $\ell^*_-(\hat{y}) = \ell^*(0, \hat{y})$. In the noiseless case, the variance is given by

$$\begin{aligned}\mathbb{V}[\ell^*(\hat{y}, Y^*)] &= \mathbb{V}\big[Y^* \ell^*_+(\hat{y}) + (1 - Y^*)\ell^*_-(\hat{y})\big] \\ &= \mathbb{V}[Y^*] \left(\ell^*_+(\hat{y}) - \ell^*_-(\hat{y})\right)^2 \end{aligned} \tag{48}$$

The unbiased estimator can be written as

$$\ell(y, \hat{y}) = \mathfrak{P}(\ell^*)(y, \hat{y}) = Y \frac{\ell^*_+(\hat{y}) + (p-1)\ell^*_-(\hat{y})}{p} + (1 - Y)\ell^*_-(\hat{y}), \tag{49}$$

with variance

$$\begin{aligned}\mathbb{V}[\ell(\hat{y}, Y)] &= \mathbb{V}\left[Y \frac{\ell^*_+(\hat{y}) + (p-1)\ell^*_-(\hat{y})}{p} + (1 - Y)\ell^*_-(\hat{y})\right] \\ &= \mathbb{V}[Y] \left(\frac{\ell^*_+(\hat{y}) + (p-1)\ell^*_-(\hat{y})}{p}\right)^2 + \mathbb{V}[Y]\,\ell^*_-(\hat{y})^2 \\ &= \mathbb{V}[Y]\,\frac{\left(\ell^*_+(\hat{y}) + (p-1)\ell^*_-(\hat{y})\right)^2 + p^2 \ell^*_-(\hat{y})^2}{p^2}. \end{aligned} \tag{50}$$

For propensities much smaller than 1, this can be approximated by (with $q := \mathbb{E}[Y] = q^* p$)

$$\approx \mathbb{V}[Y]\,\frac{\left(\ell^*_+(\hat{y}) - \ell^*_-(\hat{y})\right)^2}{p^2} = q(1-q)\frac{\left(\ell^*_+(\hat{y}) - \ell^*_-(\hat{y})\right)^2}{p^2}. \tag{51}$$

Therefore, we can calculate the ratio (using $q^* p \ll 1$)

$$\frac{\mathbb{V}[\ell^*(\hat{y}, Y^*)]}{\mathbb{V}[\ell(\hat{y}, Y)]} = \frac{(q^*(1-q^*))p^2}{q(1-q)} = \frac{(q^*(1-q^*))p^2}{q^* p(1 - q^* p)} = \frac{(1-q^*)p}{1 - q^* p} \approx (1-q^*)p. \tag{52}$$

□

Theorem 3 (Uniqueness). *Let $\ell, \ell' : \{0,1\}^l \times \mathcal{X} \longrightarrow \mathbb{R}$ denote two unbiased estimators such that*

$$\mathbb{E}[\ell^*(\mathbf{Y}^*, X)] = \mathbb{E}[\ell(\mathbf{Y}, X)] = \mathbb{E}[\ell'(\mathbf{Y}, X)] \tag{53}$$

for all distributions of $\mathbf{Y}$, $\mathbf{Y}^*$ and X that fulfill the conditions of Theorem 1 for a given vector of propensities $\mathbf{p}$. Then $\ell = \ell'$.

Proof. As the expectations need to be equal for all distributions, we can in particular choose a distribution in which $\mathbf{Y}^*$ and features X are concentrated on a single point $\mathbf{y}^*$ and x. Then

$$\mathbb{E}[\ell^*(\mathbf{Y}^*, X)] = \ell^*(\mathbf{y}^*, x). \tag{54}$$

For $\mathbf{y}^* = \mathbf{0}$ this results in the following conditions

$$\ell'(\mathbf{0}, x) = \mathbb{E}[\ell'(\mathbf{Y}, X)] = \ell^*(\mathbf{0}, x) = \mathbb{E}[\ell(\mathbf{Y}, X)] = \ell(\mathbf{0}, x). \tag{55}$$

Now we can perform the following induction: assume that for all $\mathbf{y} \in \{0,1\}^l$ with $\|\mathbf{y}\|_1 \leq k$ we have already shown that $\ell(\mathbf{y}, x) = \ell'(\mathbf{y}, x)$. Let now $\mathbf{y} \in \{0,1\}^l$ have $k+1$ positive labels, $\|\mathbf{y}\|_1 = k+1$. Then we can write

$$\begin{aligned}0 = \mathbb{E}[\ell(\mathbf{y}, x) - \ell'(\mathbf{y}, x)] &= \sum_{\mathbf{y}' \preceq \mathbf{y}} \mathbb{P}\{\mathbf{Y} = \mathbf{y}'\}\left(\ell(\mathbf{y}', x) - \ell'(\mathbf{y}', x)\right) \\ &= \mathbb{P}\{Y = \mathbf{y}\}\left(\ell(\mathbf{y}, x) - \ell'(\mathbf{y}, x)\right). \end{aligned} \tag{56}$$

Here, all the terms in the sum for which $\mathbf{y}' \prec \mathbf{y}$ are zero because of the induction Because we assume that all propensities are larger than zero, $\mathbb{P}\{Y = \mathbf{y}\} > 0$ which implies

$$\ell(\mathbf{y}, x) = \ell'(\mathbf{y}, x). \tag{57}$$

Thus, by induction $\ell = \ell'$. □

A.4 Generalization Bound

In this section we present a generalization bound. To that end, we first proof some helper results. These results may look simple and standard, but given that there exists a (peer-reviewed) published variation of the generalization bound for the noisy case (discussed below) that is incorrect, so we think it prudent to give a detailed proof here.

Lemma 1 (Lipschitz Constant)**.** *In the binary case, $\mathcal{Y} = \{0, 1\}$, let $\ell^* : \mathcal{Y} \times \hat{\mathcal{Y}} \longrightarrow [a, b]$ be a bounded, ρ-Lipschitz continuous (in the second argument) function. Then the unbiased version $\ell := \mathfrak{P}\ell^*$ has Lipschitz constant $\frac{4-2p}{p}$ and range*

$$\ell(y, \hat{y}) \in \left[\frac{a + (p-1)b}{p}, \frac{b + (p-1)a}{p}\right]. \tag{58}$$

Proof. First, we determine the Lipschitz-constant of ℓ. For $y = 0$, it is the same as that of ℓ^*, so we only need to consider the $y = 1$ case.

$$\begin{aligned}|\ell(1, \hat{y}_1) - \ell(1, \hat{y}_2)| &= \left|\frac{\ell^*(1, \hat{y}_1) + (p-1)\ell^*(0, \hat{y}_1) - \ell^*(1, \hat{y}_2) - (p-1)\ell^*(0, \hat{y}_2)}{p}\right| &(59)\\ &\le \frac{1}{p}\left|\ell^*(1, \hat{y}_1) - \ell^*(1, \hat{y}_2)\right| + \frac{1-p}{p}\left|\ell^*(0, \hat{y}_1) - \ell^*(0, \hat{y}_2)\right| &(60)\\ &\le \left(\frac{1}{p} + \frac{1-p}{p}\right)\rho\|\hat{y}_1 - \hat{y}_2\|. &(61)\end{aligned}$$

In (60) we used $0 < p \le 1$. This also implies that $\frac{2-p}{p} \ge 1$, and thus the Lipschitz constant of ℓ is given by $\frac{2-p}{p}\rho$.

Now we calculate the range of ℓ. By plugging in the largest and smallest values, we have $\forall \hat{y} \in \hat{\mathcal{Y}}$

$$\tilde{a} := \frac{a + (p-1)b}{p} \le \ell(1, \hat{y}) \le \frac{b + (p-1)a}{p} =: \tilde{b}. \tag{62}$$

Because $p \in (0, 1]$ and $a \le b$, it follows that $p - 1 < 0$ and thus

$$(p-1)b \le (p-1)a \Rightarrow a + (p-1)b \le pa, \tag{63}$$

which implies $\tilde{a} \le \ell(0, \hat{y})$. Using the analog argument for $\tilde{b}$ shows the claim. □

Lemma 2. *Let $\mathcal{H} \subset \left\{h : \mathcal{X} \longrightarrow \hat{\mathcal{Y}}\right\}$ be a function class, and $\ell^* : \mathcal{Y} \times \hat{\mathcal{Y}} \longrightarrow [a, b]$ be a bounded, ρ-Lipschitz continuous (in the second argument) function. Denote $\ell := \mathfrak{P}\ell^*$. Given a sample of n noisy training points, it holds with probability of at least $1 - \delta$ that*

$$\sup_{h \in \mathcal{H}}\left(\hat{\mathrm{R}}_\ell[h] - \mathrm{R}_\ell[h]\right) \le \frac{4-2p}{p}\rho\,\mathfrak{R}_n(\mathcal{H}) + \frac{(2-p)(b-a)}{p}\sqrt{\frac{\log(1/\delta)}{2n}} \tag{64}$$

$$\sup_{h \in \mathcal{H}}\left(\mathrm{R}_\ell[h] - \hat{\mathrm{R}}_\ell[h]\right) \le \frac{4-2p}{p}\rho\,\mathfrak{R}_n(\mathcal{H}) + \frac{(2-p)(b-a)}{p}\sqrt{\frac{\log(1/\delta)}{2n}}. \tag{65}$$

Proof. Using the notation of the proof of 1, define $c := \tilde{b} - \tilde{a}$ and $\ell_{01} : \mathcal{Y} \times \hat{\mathcal{Y}} \longrightarrow [0, 1]$ by the affine transformation $\ell_{01} = c^{-1}(\ell - \tilde{a})$ such that

$$\mathrm{R}_\ell[h] - \hat{\mathrm{R}}_\ell[h] = c\left(\mathrm{R}_{\ell_{01}}[h] - \hat{\mathrm{R}}_{\ell_{01}}[h]\right). \tag{66}$$

The right hand side can be bounded with probability $1 - \delta$ using Mohri et al. [27, Theorem 3.3] by

$$\mathrm{R}_{\ell_{01}}[h] - \hat{\mathrm{R}}_{\ell_{01}}[h] \le \mathfrak{R}_n(\ell_{01} \circ \mathcal{H}) + \sqrt{\frac{\log(1/\delta)}{2n}}. \tag{67}$$

As the Lipschitz-constant of ℓ_{01} is $c^{-1}p^{-1}(2-p)$, by the contraction lemma [36, Lemma 26.9] we have

$$\mathfrak{R}_n(\ell_{01} \circ \mathcal{H}) \le c^{-1}\frac{2-p}{p}\rho\,\mathfrak{R}_n(\mathcal{H}). \tag{68}$$

Thus with probability $1-\delta$ and $\forall h \in \mathcal{H}$

$$\mathrm{R}_\ell[h] - \hat{\mathrm{R}}_\ell[h] \le cc^{-1}\frac{2-p}{p}\rho\,\mathfrak{R}_n(\mathcal{H}) + c\sqrt{\frac{\log(1/\delta)}{2n}} \tag{69}$$

$$= \frac{2-p}{p}\rho\,\mathfrak{R}_n(\mathcal{H}) + \frac{(2-p)(b-a)}{p}\sqrt{\frac{\log(1/\delta)}{2n}} \tag{70}$$

The second bound follows by replacing f with $-f$. □

This result is very similar to Natarajan et al. [28, Lemma 8]. However, that theorem is missing a scaling factor with the range of the loss function, as argued below. For reference, the original statement of the theorem is

Theorem 10 (Natarajan et al. [28, Lemma 8])**.** *Let $l(t, y)$ be L-Lipschitz in t (for every y). Then, for any $\alpha \in (0,1)$, with probability at least $1-\delta$,*

$$\max_{f\in\mathcal{F}} \left|\hat{R}_{\tilde{l}_\alpha}(f) - R_{\tilde{l}_\alpha, D_\rho}(f)\right| \le 2L_\rho\,\mathfrak{R}_n(\mathcal{F}) + \sqrt{\frac{\log(1/\delta)}{2n}}, \tag{71}$$

where $L_\rho \le 2L/(1-\rho_{+1}-\rho_{-1})$ is the Lipschitz constant of $\tilde{l}_\alpha$.

In the first step of the proof, they invoke a "Basic Rademacher bound between risks and empirical risks" that states

$$\max_{f\in\mathcal{F}} \left|\hat{R}_{\tilde{l}_\alpha}(f) - R_{\tilde{l}_\alpha, D_\rho}(f)\right| \le 2\,\mathfrak{R}_n(\tilde{l}_\alpha \circ \mathcal{F}) + \sqrt{\frac{\log(1/\delta)}{2n}} \tag{72}$$

However, such a bound either requires the range of $\tilde{l}_\alpha$ to be a subset of $[0,1]$ [27, Thm 3.3], or introduces an additional factor in front of the square root term as in Shalev-Shwartz and Ben-David [36, Thm 26.5]. Also, they are using a two-sided bound instead of a one-sided one as in the two cited theorems, which means that δ needs to be replaced with $\delta/2$ because the square-root term comes from an application of Mc. Diamids inequality.

Theorem 11 (Generalization bound)**.** *Let $\mathcal{H}$ be a function class with Rademacher complexity $\mathfrak{R}_n(\mathcal{H})$. Let $\ell^* : \{0,1\} \times \hat{\mathcal{Y}} \longrightarrow [a,b]$ for $a < b \in \mathbb{R}$ be a binary loss function that is $\rho-$Lipschitz continuous in its second argument. Let ℓ be the corresponding unbiased estimate and denote the true risk on clean data with $\mathrm{R}^*_{\ell^*}$, and the corresponding empirical risk on noisy data as $\hat{\mathrm{R}}_\ell$, defined as follows:*

$$\mathrm{R}^*_{\ell^*}[h] := \mathbb{E}[\ell^*(\mathbf{Y}^*, h(X))]\,, \quad \hat{\mathrm{R}}_\ell[h] := \frac{1}{n}\sum_{i=1}^{n} \ell(\mathbf{y}_i, h(x_i)). \tag{73}$$

The risks are minimized at

$$\hat{r} := \inf_{h\in\mathcal{H}} \hat{\mathrm{R}}_\ell[h]\,, \quad r^\star := \inf_{h\in\mathcal{H}} \mathrm{R}^*_{\ell^*}[h]\,. \tag{74}$$

For a given sample of n points, let $\hat{h}_i$ be a sequence of classifiers such that $\hat{\mathrm{R}}_\ell\left[\hat{h}_i\right] \to \hat{r}$. Then with probability at least $1-\delta$ it holds that

$$\lim_{i\to\infty} \mathrm{R}^*_{\ell^*}\left[\hat{h}_i\right] \le \inf_{h\in\mathcal{H}} \mathrm{R}^*_{\ell^*}[h] + 2\frac{2-p}{p}\left(\rho\,\mathfrak{R}_n(\mathcal{H}) + (b-a)\sqrt{\frac{\log(2/\delta)}{2n}}\right) \tag{75}$$

where $\mathrm{R}^*_{\ell^*}$ denotes the true risk on clean data, and $\hat{\mathrm{R}}_\ell$ denotes the empirical risk on observed data.

Proof. Let $\epsilon > 0$, then there exists $h' \in \mathcal{H}$ and $k \in \mathbb{N}$ such that $\forall i \ge k$

$$r' := \mathrm{R}^*_{\ell^*}[h'] \le r^\star + \epsilon, \quad \hat{r}_i := \hat{\mathrm{R}}_\ell\left[\hat{h}_i\right] \le \hat{r} + \epsilon. \tag{76}$$

Due to the optimality of $\hat{r}$ it holds that

$$\hat{\mathrm{R}}_\ell[h'] \ge \hat{r} \ge \hat{r}_i - \epsilon \;\Rightarrow\; \hat{\mathrm{R}}_\ell\left[\hat{h}_i\right] - \hat{\mathrm{R}}_\ell[h'] \le \epsilon. \tag{77}$$

We can apply Lemma 2 to the function class $\{h'\}$ using that $\mathfrak{R}_n(\{h' = 0\}) = 0$ to get with probability $1 - \delta/2$

$$\hat{\mathrm{R}}_\ell[h'] - \mathrm{R}_\ell[h'] \leq \frac{(2-p)(b-a)}{p}\sqrt{\frac{\log(2/\delta)}{2n}}. \tag{78}$$

Using unbiasedness of ℓ and the near optimality (77) of $\hat{h}_i$ regarding $\hat{\mathrm{R}}_\ell$ to bound

$$\begin{aligned}
\mathrm{R}^*_{\ell*}\left[\hat{h}_i\right] - \mathrm{R}^*_{\ell*}[h'] &= \mathrm{R}_\ell\left[\hat{h}_i\right] - \mathrm{R}_\ell[h'] && \text{(unbiasedness)}\\
&= \mathrm{R}_\ell\left[\hat{h}\right] - \hat{\mathrm{R}}_\ell\left[\hat{h}\right] + \hat{\mathrm{R}}_\ell\left[\hat{h}\right] - \hat{\mathrm{R}}_\ell[h'] + \hat{\mathrm{R}}_\ell[h'] - \mathrm{R}_\ell[h'] \\
&\leq \mathrm{R}_\ell\left[\hat{h}\right] - \hat{\mathrm{R}}_\ell\left[\hat{h}\right] + \epsilon + \hat{\mathrm{R}}_\ell[h'] - \mathrm{R}_\ell[h'] && \text{(optimality)}\\
&\leq \sup_{h\in\mathcal{H}}\left(\mathrm{R}_\ell[h] - \hat{\mathrm{R}}_\ell[h]\right) + \hat{\mathrm{R}}_\ell[h'] - \mathrm{R}_\ell[h'] + \epsilon.
\end{aligned}$$

Applying a union bound to the remaining two terms, with probability $1 - \delta$

$$\begin{aligned}
\mathrm{R}^*_{\ell*}\left[\hat{h}_i\right] - r' &\leq \epsilon + \frac{4-2p}{p}\mathfrak{R}(\mathcal{H}) + \frac{(2-p)(b-a)}{p}\sqrt{\frac{\log(2/\delta)}{2n}} + \frac{(2-p)(b-a)}{p}\sqrt{\frac{\log(2/\delta)}{2n}} \\
&= \epsilon + \frac{4-2p}{p}\left(\mathfrak{R}_n(\mathcal{H}) + (b-a)\sqrt{\frac{\log(2/\delta)}{2n}}\right).
\end{aligned} \tag{79}$$

Substituting $r^\star$ yields

$$\mathrm{R}^*_{\ell*}\left[\hat{h}_i\right] - r^\star \leq 2\epsilon + \frac{4-2p}{p}\left(\mathfrak{R}_n(\mathcal{H}) + (b-a)\sqrt{\frac{\log(2/\delta)}{2n}}\right). \tag{80}$$

Now we cen let $i \to \infty$ and then $\epsilon \to 0$ to prove the claim. □

B Additional Details for the Experiments

B.1 Data Generation

We generated the semi-artificial dataset as follows. As a basis, we used the AmazonCat-13k [23] dataset, which has $13\,000$ labels, $n = 1\,200\,000$ training instances and $300\,000$ test instances. We sorted the labels by their frequency, and then removed all labels except those among the top-100 most frequent ones. This gives us a new training set $\mathcal{D} = \left\{(x_i, \mathbf{y}_i) : i \in [n], \mathbf{y}_i \in \{0,1\}^{100}\right\}$.

We then artificially remove labels. To that end, we generate a random variable $\mathbf{M}$ such that $\mathbb{P}\{M_j = 1\} = \frac{1}{2+18\cdot j/100}$, i.e. the inverse propensity increases linearly from 2 to 20. We further split the training data into a training set and a validation set. Let σ be a permutation of $[n]$, leading to a random shuffle of the data, then this leads to the datasets

$$\mathcal{D}_{\text{train}} = \left\{(x_{\sigma(i)}, \mathbf{y}_{\sigma(i)} \odot \mathbf{M}) : 0 \leq i \leq 0.7 \cdot n\right\} \tag{81}$$

$$\mathcal{D}_{\text{val}} = \left\{(x_{\sigma(i)}, \mathbf{y}_{\sigma(i)} \odot \mathbf{M}) : 0.7n < i \leq n\right\} \tag{82}$$

We repeated this process 5 times to get different variations of the dataset in order to be able to estimate the standard deviation marked in the plots. For each sweep over the regularization parameter, we used a fixed version of the dataset in order to be able to meaningfully determine the optimal regularization parameter as one would do in a real-data scenario. This optimal parameter is then used to calculate the actual performance on the test set, as presented in Table 2.

B.2 Training Details

The network is optimized using Adam [21] with an initial learning rate of 10^{-4} for the first 15 epochs and 10^{-5} for the remaining five epochs, with a mini-batch size of 512. The learning rate was chosen this low in order to ensure stable training even for the unbiased loss function for the normalized reductions, which can otherwise become problematic due to their large variance.

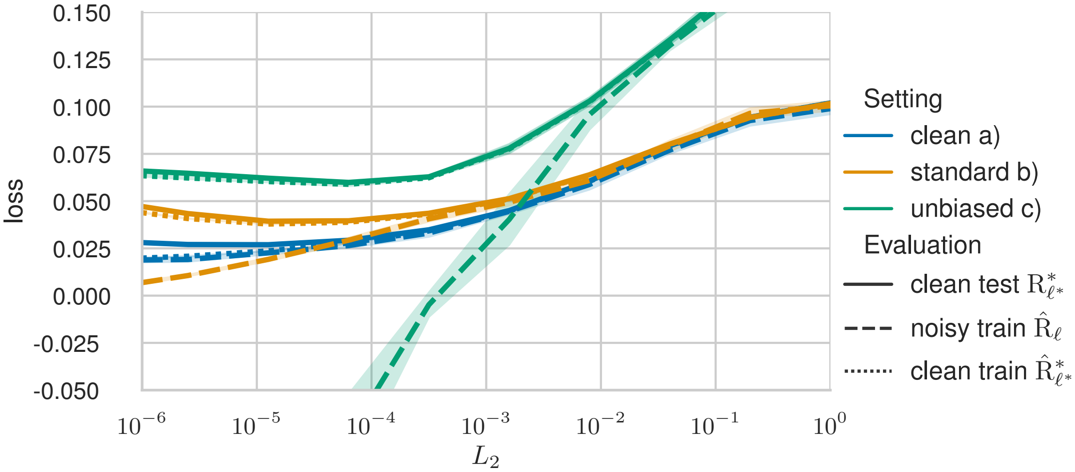

Figure 3: Normalized binary cross-entropy for different regularization strengths, evaluated on noisy training data, clean training data, and clean test data. In this setting, the unbiased estimate is not useful for training, and the standard loss function results in better performance.

By training with the standard loss, we get an upper bound how well the given network could do without noise (config **a**) and how poorly it would perform without any mitigations (config **b**). In the first setting, we expect the calculated loss on clean and noisy data to match, since the model cannot overfit to the specific noise pattern in the training data, whereas in configurations **b)** to **d)** we expect to see a difference. As we remove labels independently of features, the unbiased estimate on noisy test data should be equal to the actual value on clean test data, if the test set contains sufficient samples to result in accurate estimates.

Even though the average number of relevant labels per instance is low, there may still be individual data points with a large enough number of relevant labels to make calculating the sum over all subsets as in Theorem 2 impractical. For that reason, instead of calculating the entire sum, we instead choose a sampling approach by taking a uniform random subsample of the summands. For the experiments presented in this paper, we used 32 samples (label subsets) for each data point. While this, in principle, may result in an increased variance, but we observed almost no change in behaviour when doubling the number of samples.

B.3 More Results with Artificial Noise

For the normalized BCE reduction, we can see (Figure 3) that the unbiased loss is not helpful, as it underperforms using the standard BCE across the entire range of tested regularization parameters, even when the regularization is so strong that only minimal overfitting happens.

The situation is similar for the normalized CCE loss (Figure 2, bottom), in that the unbiased estimate results in bad test performance. At strong regularization, it performs slightly better than the standard loss, but its minimum is much higher. Contrary to the normalized OVA case, in this situation we also have an upper-bound available, which turns out to perform slightly worse than the standard loss for low regularization, but significantly better for strong regularization, with an overall improved minimum.

In the normalized setting, the variance of the unbiased estimate becomes already noticeable in the evaluation procedure. Even when training with full labels (blue curves), there is a slight difference between evaluating the loss function on clean data (dotted) and its unbiased estimate on noisy data (dashed). This variance also expresses itself in the error intervals for the unbiased training, which show that multiple runs can lead to significantly different outcomes.

For the decomposable CCE loss Figure 4, upper-bound and unbiased loss are the same function. Again, it results in worse test loss for weak regularization, but better test loss for strong regularization.

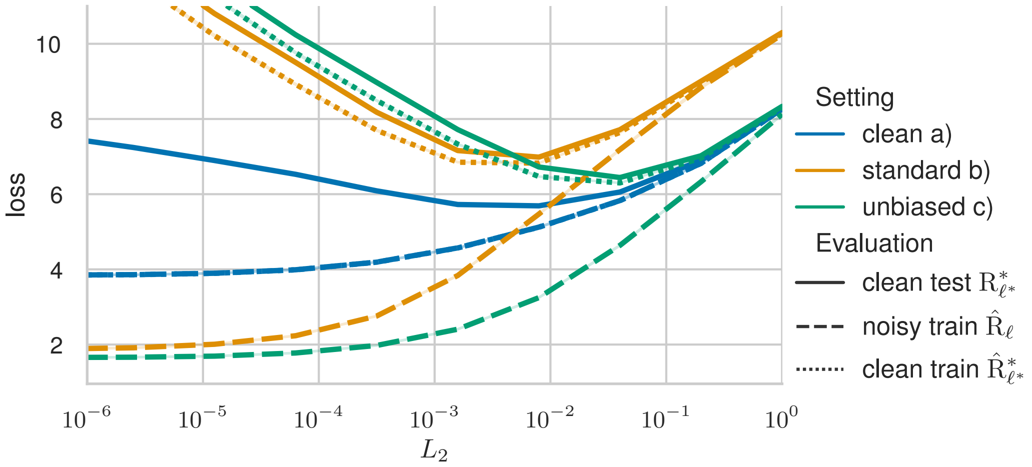

Figure 4: Categorical cross-entropy for different regularization strengths, evaluated on noisy training data, clean training data, and clean test data. In this setting the unbiased estimate is already a convex function, so no separate bound is used.

Overall, its minimum is at stronger regularization than that of standard loss, and results in lower test loss.

Table 2: Training results (metrics evaluated on clean test data) on modified AmazonCat-13K data for using different loss functions in their **S**tandard, **U**nbiased or Upper-**B**ounded variants. BCE denotes the binary cross-entropy corresponding to a OvA decomposition, and CCE denotes (softmax) categorical cross-entropy corresponding to a PaL decomposition. Reference runs with clean training data are shown in gray. Bold fonts denote the overall best result (on noisy training data), underlined means best among the basic loss function.

Setting		Precision			Recall			Loss	Reg.
		P@1	P@3	P@5	R@1	R@3	R@5		
S-BCE	a)	91.7	61.2	43.3	50.0	81.0	88.9	$4.1 \cdot 10^{-2}$	$1.26 \cdot 10^{-5}$
S-BCE	b)	83.9	53.9	38.3	43.8	71.7	80.2	$8.9 \cdot 10^{-2}$	$6.31 \cdot 10^{-5}$
U-BCE	c)	**86.9**	**57.5**	**40.9**	$\underline{46.2}$	**76.1**	**84.5**	$\underline{5.8 \cdot 10^{-2}}$	$3.16 \cdot 10^{-4}$
B-BCE	d)	86.4	57.1	40.7	46.1	75.9	84.4	$6.3 \cdot 10^{-2}$	$3.16 \cdot 10^{-4}$
S-NBCE	a)	87.7	58.7	41.9	48.6	79.1	87.3	$2.8 \cdot 10^{-2}$	$4.73 \cdot 10^{-6}$
S-NBCE	b)	$\underline{83.6}$	$\underline{53.6}$	$\underline{38.1}$	$\underline{43.9}$	$\underline{71.9}$	$\underline{80.3}$	$\underline{3.9 \cdot 10^{-2}}$	$2.1 \cdot 10^{-5}$
U-NBCE	c)	72.5	45.3	32.3	40.5	64.5	72.3	$6.2 \cdot 10^{-2}$	$3.41 \cdot 10^{-5}$
B-NBCE	d)	75.8	45.9	32.1	37.1	60.0	67.5	$5.6 \cdot 10^{-2}$	$3.4 \cdot 10^{-4}$
S-CCE	a)	89.4	59.9	42.8	48.8	79.9	88.2	5.7	$7.94 \cdot 10^{-3}$
S-CCE	b)	85.0	53.9	38.3	44.8	72.2	80.5	7	$7.94 \cdot 10^{-3}$
B-CCE	c,d)	$\underline{86.4}$	$\underline{56.9}$	$\underline{40.6}$	$\underline{46.3}$	$\underline{76.0}$	**84.5**	$\underline{6.4}$	$3.98 \cdot 10^{-2}$
S-NCCE	a)	88.9	59.8	42.6	49.0	80.2	88.5	1.6	$4.43 \cdot 10^{-4}$
S-NCCE	b)	85.6	54.9	39.1	45.7	74.1	82.5	2.1	$2.06 \cdot 10^{-3}$
U-NCCE	c)	71.9	43.8	30.9	40.6	64.3	71.7	3.3	0.13
B-NCCE	d)	$\underline{86.7}$	$\underline{56.3}$	$\underline{40.0}$	**46.7**	$\underline{75.9}$	$\underline{84.1}$	$\underline{1.9}$	$2.39 \cdot 10^{-2}$

B.4 Datasets with Approximate Propensity Specification

In Jain et al. [16], the authors developed an empirical model to estimate propensity values for different labels in extreme classification datasets. This misspecification can lead to unbiased estimates that are far outside the range of credible values — precision at k with values larger than 100%. This is

Table 3: Unbiased and vanilla test precision, as well as vanilla recall, on eurlex data. Bold font marks the overall best result for a metric, underlining signifies the best variation among a given loss type.

Setting		Unbiased			Biased			Biased			
		P@1	P@3	P@5	P@1	P@3	P@5	R@1	R@3	R@5	Reg
V-BCE	b)	162.3	140.3	119.2	76.9	64.0	52.9	15.6	38.1	51.5	$3.16 \cdot 10^{-6}$
U-BCE	c)	**181.5**	**151.0**	125.0	75.2	59.4	48.8	15.3	35.2	47.4	$3.16 \cdot 10^{-7}$
B-BCE	d)	172.1	147.7	**126.1**	**78.1**	**64.2**	**53.6**	**15.9**	**38.2**	**52.2**	$3.16 \cdot 10^{-6}$
V-NBCE	b)	145.4	124.1	105.6	68.6	57.1	47.8	14.1	34.1	46.7	$1 \cdot 10^{-6}$
B-NBCE	c)	142.7	116.4	98.6	63.5	52.5	44.4	13.0	31.3	43.4	$3.16 \cdot 10^{-7}$
V-CCE	b)	142.6	126.7	109.0	65.9	57.1	48.0	13.4	33.9	46.8	$1 \cdot 10^{-2}$
U-CCE	c,d)	170.8	137.6	115.3	69.9	57.6	48.5	14.2	34.1	47.2	0.1
V-NCCE	b)	146.2	125.9	108.7	67.7	57.2	48.3	13.8	34.1	47.1	$3.16 \cdot 10^{-3}$
U-NCCE	c)	2.5	3.8	3.8	1.2	1.7	1.6	0.4	1.1	1.7	100
B-NCCE	d)	170.4	136.7	114.7	67.2	56.5	47.9	13.7	33.7	46.8	$3.16 \cdot 10^{-3}$

Table 4: Unbiased and vanilla test precision, as well as vanilla recall, on AmazonCat data.

Setting		Unbiased			Biased			Biased			
		P@1	P@3	P@5	P@1	P@3	P@5	R@1	R@3	R@5	Reg
S-BCE	b)	93.4	75.2	59.5	88.6	68.5	52.7	25.1	52.4	63.2	$1 \cdot 10^{-10}$
U-BCE	c)	95.1	78.4	63.2	87.1	69.1	54.1	24.6	52.6	64.6	$1 \cdot 10^{-11}$
B-BCE	d)	97.8	83.1	68.8	**89.4**	71.7	56.9	**25.2**	54.5	67.6	$1 \cdot 10^{-11}$
S-NBCE	b)	82.6	54.4	39.9	79.3	51.1	37.1	22.9	42.1	48.7	$1 \cdot 10^{-11}$
B-NBCE	d)	75.1	49.1	36.0	72.2	46.3	33.7	21.2	39.1	45.3	$1 \cdot 10^{-11}$
S-CCE	b)	102.5	90.2	76.4	86.2	73.8	60.3	24.1	56.5	71.8	$1 \cdot 10^{-3}$
U-CCE	c,d)	114.4	96.1	79.3	85.4	73.4	60.4	23.8	56.2	71.8	$1 \cdot 10^{-2}$
S-NCCE	b)	106.3	92.2	77.4	88.2	**74.7**	**60.7**	25.0	**57.5**	**72.6**	$1 \cdot 10^{-6}$
B-NCCE	d)	**119.4**	**97.7**	**79.6**	86.7	73.8	60.4	24.6	56.9	72.3	$1 \cdot 10^{-4}$

masked by current XMC papers typically reporting a normalized version of these estimates [35], dividing by their largest possible value, see for example Jain et al. [16, Section 7].

However, we can use these datasets to demonstrate that the approaches proposed in this paper lead to relative improvements, even if the absolute numbers are not meaningful.

Eurlex For the Eurlex dataset, we did not run the very unstable setting of training with unbiased normalized BCE loss. Furthermore, the much smaller size of the dataset compared to AmazonCat means that more epochs are necessary in order to ensure sufficiently many weight updates to minimize the empirical risk. Therefore, we used an initial learning rate of 10^{-3} for the first 60 epochs and 10^{-4} for the remaining 20 epochs.

Because we know that the assumed missing label model is misspecified to some degree, we used the unbiased estimate of P@3 on validation data as the criterion for selecting the best regularization parameter. This quantity is more robust to overfitting than the loss function itself.

The variance when using the unbiased estimate for normalized PAL reduction is so large that training becomes impossible. Even after decreasing the learning rate by a factor of 100 and increasing the regularization to $\lambda = 100$ (a factor of 1000 larger than the optimal regularization for the non-normalized case), the validation loss remained wildly fluctuating between -1×10^6 and 1×10^6.

The results at the optimal regularization determined in this way are summarized in Table 3. Strikingly, the upper-bound BCE loss results in improvements both in the unbiased estimate and in the biased (vanilla) metric. We conjecture that this is because the weighting introduced in this case is similar to re-weighting the data to address the imbalance in the label distribution.

Table 5: Unbiased and vanilla test precision, as well as vanilla recall, on Wiki10-31K data.

Setting		Unbiased			Biased			Biased			
		P@1	P@3	P@5	P@1	P@3	P@5	R@1	R@3	R@5	Reg
S-BCE	b)	108.1	95.4	85.6	83.6	69.0	59.7	4.9	12.0	16.9	$2.8 \cdot 10^{-9}$
U-BCE	c)	108.8	85.3	71.7	36.1	26.0	21.3	2.2	4.6	6.2	$2.8 \cdot 10^{-9}$
B-BCE	d)	**133.2**	**121.6**	**109.1**	**84.7**	**72.1**	**62.7**	**5.0**	**12.6**	**17.9**	$5.25 \cdot 10^{-8}$
S-NBCE	b)	$\underline{87.5}$	$\underline{58.7}$	45.7	80.4	$\underline{52.4}$	$\underline{40.0}$	$\underline{4.7}$	$\underline{8.9}$	$\underline{11.1}$	$1 \cdot 10^{-8}$
U-NBCE	c)	0.3	0.3	0.3	0.1	0.1	0.1	0.0	0.0	0.0	$1 \cdot 10^{-6}$
B-NBCE	d)	$\underline{87.5}$	55.9	41.2	$\underline{80.8}$	50.6	36.7	$\underline{4.7}$	8.5	10.1	$1 \cdot 10^{-7}$
S-CCE	b)	96.2	81.1	73.5	$\underline{73.7}$	$\underline{59.1}$	$\underline{51.7}$	$\underline{4.3}$	$\underline{10.1}$	$\underline{14.5}$	5.5
U-CCE	c,d)	$\underline{109.4}$	$\underline{94.0}$	$\underline{85.3}$	70.7	57.5	50.5	4.1	9.8	14.2	7.75
S-NCCE	b)	100.1	85.2	$\underline{76.9}$	75.7	61.5	53.5	$\underline{4.4}$	$\underline{10.6}$	15.1	0.1
B-NCCE	d)	$\underline{116.6}$	$\underline{99.8}$	$\underline{90.1}$	$\underline{76.1}$	$\underline{61.7}$	$\underline{53.8}$	$\underline{4.4}$	$\underline{10.6}$	$\underline{15.2}$	0.1

AmazonCat For the full AmazonCat data, in order to keep memory consumption low enough so that the model could be trained on a GPU, we parametrized the model with two linear layers such that there is a hidden representation with 512 units. This effectively limits the rank of the linear model to 512, thus potentially reducing its expressiveness. However, our goal here is to show how the different loss functions interact with missing labels, not to produce state-of-the-art results. We used a batch size of 256 and a learning rate of 5×10^{-4} for the first fifteen epochs and 5×10^{-5} for the remaining five epochs.

In contrast to the eurlex dataset, where best results are achieved using OVA reduction, for AmazonCat the best precisiopn results are based on normalized PAL. This seems to go against the results of Menon et al. [25], which state that non-normalized losses should be used for optimality in precision at k. However, the results presented there concern the asymptotic case of infinite data, and thus need not necessarily apply when training from a finite dataset.

Wiki10 For the Wiki10 dataset, we use the same training setup as for AmazonCat, with a learning rate schedule as for the Eurlex data, i.e. 60 epochs with 10^{-3} and 20 epochs with 10^{-4}. The results are given in Table 5. As in the Eurlex case, we observe that the OVA reduction results in the best performance.

C Multilabel Reductions

In large-scale multilabel learning are often formulated as (bipartite) ranking tasks where one is mostly interested in the prediction at the top. For example, in related product recommendation, a website may have three slots to display related products. The system ranks all products according to their relatedness, and takes the three most related ones. In this ranking, it is very important that the three top ranks are relevant to the reference product, whereas the relative ordering of recommendations at ranks 1000 and 2000 is unimportant.

As a consequence, typical performance measures in XMLC are calculated "at k", which means that we transform the ranking into a binary prediction by taking the top-k ranked items to be positive, for some fixed number k. This is in contrast to the more common classification prediction where a positive prediction is determined by a fixed thresholds (e.g. 0.5 if the scoring represents probabilities), and the number of positive predictions can vary across instances.

This leads to the performance metrics precision-at-k and recall-at-k, defined through

$$\mathrm{P@k}(\mathbf{y}, \hat{\mathbf{y}}) = k^{-1} \sum_{j \in \mathrm{top_k}(\hat{\mathbf{y}})} y_j, \quad \mathrm{R@k}(\mathbf{y}, \hat{\mathbf{y}}) = \|\mathbf{y}\|_1^{-1} \sum_{j \in \mathrm{top_k}(\hat{\mathbf{y}})} y_j. \tag{83}$$

These objectives are nondifferentiable and thus difficult to optimize. A common strategy to tackle this problem is to use a surrogate loss that is differentiable but, in the infinite-data limit, still yields the correct optimizer. Such surrogates can be constructed by *multilabel reductions* as laid out in Menon et al. [25].

Table 6: Decompositions of Multilabel Reductions

Name	Reduction	f	g_j	z_j
OvA	Binary	$\sum_{j=1}^{l} \ell_{\text{BC}}(0, \hat{y}_j)$	$\ell_{\text{BC}}(1, \hat{y}_j) - \ell_{\text{BC}}(0, \hat{y}_j)$	y_j
OvA-N	Binary	$\sum_{j=1}^{l} \ell_{\text{BC}}(0, \hat{y}_j)$	$\ell_{\text{BC}}(1, \hat{y}_j) - \ell_{\text{BC}}(0, \hat{y}_j)$	$y_j/(\sum_{i\neq j} y_i)$
PaL	Multiclass	0	$\ell_{\text{MC}}(j, \hat{\mathbf{y}})$	y_j
PaL-N	Multiclass	0	$\ell_{\text{MC}}(j, \hat{\mathbf{y}})$	$y_j/(\sum_{i\neq j} y_i)$

For $\ell_{\text{BC}} : \{0, 1\} \times \mathbb{R} \longrightarrow \mathbb{R}$ a binary loss and $\ell_{\text{MC}} : [l] \times \mathbb{R}^l \longrightarrow \mathbb{R}$ a multiclass loss function, the *one versus all* and the *pick all labels* reduction are given by

$$\ell^*_{\text{OvA}}(\mathbf{y}, \hat{\mathbf{y}}) = \sum_{j=1}^{l} \ell_{\text{BC}}(y_j, \hat{y}_j) \qquad = \sum_{j=1}^{l} y_j \left(\ell_{\text{BC}}(1, \hat{y}_j) - \ell_{\text{BC}}(0, \hat{y}_j)\right) + \ell_{\text{BC}}(0, \hat{y}_j) \quad \text{(One vs All)}$$

$$\ell^*_{\text{PAL}}(\mathbf{y}, \hat{\mathbf{y}}) = \sum_{j:y_j=1} \ell_{\text{MC}}(j, \hat{\mathbf{y}}) \qquad = \sum_{j=1}^{l} y_j \ell_{\text{MC}}(j, \hat{\mathbf{y}}). \quad \text{(Pick all Labels)}$$

By replacing the label y_j with a normalized version $y_j/\|\mathbf{y}\|_1$, this gives the corresponding normalized reductions

$$\ell^*_{\text{OvA-N}}(\mathbf{y}, \hat{\mathbf{y}}) = \sum_{j=1}^{l} \frac{y_j}{\|\mathbf{y}\|_1} \left(\ell_{\text{BC}}(1, \hat{y}_j) - \ell_{\text{BC}}(0, \hat{y}_j)\right) + \ell_{\text{BC}}(0, \hat{y}_j) \tag{84}$$

$$\ell^*_{\text{PAL-N}}(\mathbf{y}, \hat{\mathbf{y}}) = \sum_{j=1}^{l} \frac{y_j}{\|\mathbf{y}\|_1} \ell_{\text{MC}}(j, \hat{\mathbf{y}}). \tag{85}$$

These reductions can be written in a common structure, decomposing into a part linear in the (normalized) label and one part independent of the label, as done in equation (4) in the main part of the paper, repeated here for convenience:

$$\ell^*(\mathbf{y}, \hat{\mathbf{y}}) = f(\hat{\mathbf{y}}) + \sum_{j=1}^{l} z_j g_j(\hat{\mathbf{y}}). \tag{4}$$

The corresponding structures for f and g are listed in Table 6.

If the underlying base losses are sufficiently well behaved, then the reductions can be used to minimize the regret also in regards to the original multiclass loss P@k or R@k

Proposition 3. *Let ℓ_{MC} be a consistent multiclass loss, and ℓ_{BC} be a λ-strong proper composite loss. Denote with*

$$\operatorname{reg}(f, \ell) := \mathbb{E}[\ell(Y^*, f(X))] - \inf_h \mathbb{E}[\ell(Y^*, h(X))] \tag{86}$$

the regret of a scorer $f : \mathcal{X} \longrightarrow [0, 1]$ measured with loss function ℓ. It holds that

$$\operatorname{reg}(f, \text{P@k}) \leq 2\sqrt{2/\lambda} \cdot \max_{j\in[l]} \sqrt{\operatorname{reg}(f_j, \ell_{BC})} \tag{87}$$

$$\operatorname{reg}(f, \text{R@k}) \leq \sqrt{2/\lambda} \cdot \operatorname{reg}(f, \ell^*_{OvA\text{-}N}) \tag{88}$$

Further, fo for a sequence of scorers $f_n : \mathcal{X} \longrightarrow [0, 1]$:

$$\operatorname{reg}(f_n, \ell^*_{PAL}) \to 0 \implies \operatorname{reg}(f_n, \text{P@k}) \to 0 \tag{89}$$

$$\operatorname{reg}(f_n, \ell^*_{PAL\text{-}N}) \to 0 \implies \operatorname{reg}(f_n, \text{R@k}) \to 0. \tag{90}$$

For more details, we refer to Wydmuch et al. [42] for the OVA case, and Menon et al. [25] for the other three cases.