Appendix to "Revisiting F-measure Optimization in Multi-Label Classification: A Sampling-based Approach"

Zixun Wang
Department of Statistics and Data Science
The Chinese University of Hong Kong
craddywang@gmail.com

A. Proof

A.1. Proof of Theorem 1

The theorem comes from Dembczynski et al. [1], we provide a proof here for completeness.

Proof. Partition the label space $\{0,1\}^q$ into $q+1$ disjoint subset $\{0,1\}^q = \bigcup_{\tau=0}^q \mathcal{Z}_\tau$, where $\mathcal{Z}_\tau = \{\mathbf{y} \in \{0,1\}^q : \|\mathbf{y}\|_1 = \tau\}$. The Bayes classifier for the F-measure is then given by:

$$\mathbf{h}^*(\mathbf{x}) = \operatorname*{argmax}_{\mathbf{z}\in\{\mathbf{z}_0^*,\cdots,\mathbf{z}_q^*\}} \mathbb{E}\Big(\mathrm{F}_1(\mathbf{Y},\mathbf{z})|\mathbf{X}=\mathbf{x}\Big) \tag{1}$$

$$\text{subject to} \quad \mathbf{z}_\tau^* = \operatorname*{argmax}_{\mathbf{z}\in\mathcal{Z}_\tau} \mathbb{E}\Big(\mathrm{F}_1(\mathbf{Y},\mathbf{z})|\mathbf{X}=\mathbf{x}\Big), \quad \text{for} \quad 0 \le \tau \le q. \tag{2}$$

By definition of F-measure and expanding the expectation over $\mathbf{Y}$, the inner optimization problem in Equation (2) can be expressed by:

$$\begin{aligned}\mathbf{z}_\tau^* &= \operatorname*{argmax}_{\mathbf{z}\in\mathcal{Z}_\tau} \sum_{\mathbf{y}\in\{0,1\}^q} \frac{2\sum_{j=1}^q z_j Y_j \mathbb{P}(\mathbf{y}|\mathbf{X}=\mathbf{x})}{\tau + \|\mathbf{y}\|_1} = \operatorname*{argmax}_{\mathbf{z}\in\mathcal{Z}_\tau} 2\sum_{j=1}^q z_j \sum_{\mathbf{y}\in\{0,1\}^q} \frac{Y_j \mathbb{P}(\mathbf{y}|\mathbf{X}=\mathbf{x})}{\tau + \|\mathbf{y}\|_1} \\ &= \operatorname*{argmax}_{\mathbf{z}\in\mathcal{Z}_\tau} 2\sum_{j=1}^q z_j \sum_{k=1}^q \frac{\mathbb{P}(y_j=1, \|\mathbf{y}\|_1 = k|\mathbf{X}=\mathbf{x})}{\tau + k} = \operatorname*{argmax}_{\mathbf{z}\in\mathcal{Z}_\tau} 2\sum_{j=1}^q z_j \Lambda_{j,\tau}(\mathbf{x}).\end{aligned}$$

The second equality is by swapping the summations, and the third equality is by partitioning $\mathbf{y} \in \{0,1\}^q$ into cases of $\|\mathbf{y}\|_1 = k$ for $1 \le k \le q$.

For $\mathbf{z} \in \mathcal{Z}_\tau$, only τ of the q labels z_j $(1 \le j \le q)$ can be non-zero in $\mathcal{Z}_\tau$. Therefore, to maximize the above expression, top-τ labels with the largest $\Lambda_{j,\tau}(\mathbf{x})$ $(1 \le j \le q)$ should be selected. This leads to:

$$\mathbf{z}_\tau^* = \operatorname*{argmax}_{\mathbf{z}\in\mathcal{Z}_\tau} 2\sum_{j=1}^q z_j \Lambda_{j,\tau}(\mathbf{x}) = \Big(\mathbb{1}(j \in \mathrm{Top}_\tau(\mathbf{\Lambda}_{1:q,\tau}(\mathbf{x})))\Big)_{j=1}^q, \tag{3}$$

$$\max_{\mathbf{z}\in\mathcal{Z}_\tau} \mathbb{E}\Big(\mathrm{F}_1(\mathbf{Y},\mathbf{z})|\mathbf{X}=\mathbf{x}\Big) = 2 \sum_{j\in\mathrm{Top}_\tau(\mathbf{\Lambda}_{1:q,\tau}(\mathbf{x}))} \Lambda_{j,\tau}(\mathbf{x}) = 2\omega_\tau(\mathbf{x}), \tag{4}$$

where $\omega_\tau(\mathbf{x})$ is defined in Theorem 1. Finally, it suffices to first determine the optimal $\tau^*(\mathbf{x})$ that maximizes the outer optimization problem in Equation (1) and then select the corresponding $\mathbf{z}^*_{\tau^*(\mathbf{x})}$ as $\mathbf{h}^*(\mathbf{x})$. This leads to:

$$\mathbf{h}^*(\mathbf{x}) = \mathbf{z}^*_{\tau^*(\mathbf{x})} = \Big(\mathbb{1}(j \in \mathrm{Top}_{\tau^*(\mathbf{x})}(\mathbf{\Lambda}_{1:q,\tau^*(\mathbf{x})}(\mathbf{x})))\Big)_{j=1}^q, \quad \text{s.t.} \quad \tau^*(\mathbf{x}) = \operatorname*{argmax}_{0\le\tau\le q} \omega_\tau(\mathbf{x}).$$

□

A.2. Proof of Theorem 2

Proof. Let $\widetilde{\mathbb{P}}$ be the learned conditional distribution and $\widetilde{V}_{j,k} = \widetilde{\mathbb{P}}(Y_j = 1, \|\mathbf{Y}\|_1 = k|\mathbf{X} = \mathbf{x})$. We decompose the error as follows, and then bound the two terms separately:

$$|\widehat{V}_{j,k} - V_{j,k}| \leq |\widehat{V}_{j,k} - \widetilde{V}_{j,k}| + |\widetilde{V}_{j,k} - V_{j,k}|.$$

Firstly, $\widehat{V}_{j,k}$ is the average of m i.i.d. bernoulli random variables, and thus by Hoeffding's inequality, we have, with probability at least $1 - \delta$:

$$|\widehat{V}_{j,k} - \widetilde{V}_{j,k}| \leq \sqrt{\frac{\log(2/\delta)}{2m}}. \tag{5}$$

Then we bound $|\widetilde{V}_{j,k} - V_{j,k}|$. The convergence rate of excess risk with respect to binary cross-entropy loss implies that:

$$\begin{aligned}\mathbb{E}\Big(\ell_{\text{BCE}}(\widehat{\psi}_j(\mathbf{X}, \mathbf{Y}_{<j}), Y_j)\Big) - \mathbb{E}\Big(\ell_{\text{BCE}}(g_j(\mathbf{X}, \mathbf{Y}_{<j}), Y_j)\Big) &= \mathbb{E}[\log g_j(\mathbf{X}, \mathbf{Y}_{<j}) - \log \widehat{\psi}_j(\mathbf{X}, \mathbf{Y}_{<j})] \\ &= \text{KL}(\widehat{\psi}_j(\mathbf{X}, \mathbf{Y}_{<j})||g_j(\mathbf{X}, \mathbf{Y}_{<j})) = O_p(\epsilon_n),\end{aligned}$$

where KL denotes the Kullback-Leibler divergence. For the joint distribution,

$$\text{KL}(\mathbb{P}||\widetilde{\mathbb{P}}) = \sum_{j=1}^{q} \mathbb{E}\Big(\ell_{\text{BCE}}(\widehat{\psi}_j(\mathbf{X}, \mathbf{Y}_{<j}), Y_j)\Big) = O_p(q\epsilon_n).$$

By Pinsker's inequality,

$$\text{TV}(\mathbb{P}, \widetilde{\mathbb{P}}) \leq \sqrt{\frac{1}{2}\text{KL}(\mathbb{P}||\widetilde{\mathbb{P}})} = O_p(\sqrt{q\epsilon_n}),$$

where TV denotes the total variation distance. Therefore,

$$|\widetilde{V}_{j,k} - V_{j,k}| \leq \text{TV}(\mathbb{P}, \widetilde{\mathbb{P}}) = O_p(\sqrt{q\epsilon_n}). \tag{6}$$

Combining Equation (5) and Equation (6), with probability at least $1 - \delta'$ and a constant C:

$$|\widehat{V}_{j,k} - V_{j,k}| \leq \sqrt{\frac{\log(2/\delta')}{2m}} + C\sqrt{q\epsilon_n}.$$

Let $\delta = \delta'/q^2$, then with probability at least $1 - \delta$ and a constant C, we have:

$$\begin{aligned}\|\widehat{V} - V\|_F &= \sqrt{\sum_{j=1}^{q}\sum_{k=1}^{q}(\widehat{V}_{j,k} - V_{j,k})^2} \leq \sqrt{q^2\Big(\sqrt{\frac{\log(2q^2/\delta)}{2m}} + C\sqrt{q\epsilon_n}\Big)^2} \\ &= q\Big(\sqrt{\frac{\log(2q^2/\delta)}{2m}} + C\sqrt{q\epsilon_n}\Big).\end{aligned}$$

□

B. Sparse Distribution in Direct Multinomial Estimation

Imbalance is an inherent issue in multi-label classification (MLC) datasets, where the number of positive samples for each label is typically much smaller than the number of negative samples. This imbalance is further exacerbated by the MN approach, resulting in a highly sparse multinomial distribution. Figure 1 shows the proportion of positive samples for each label across the eight datasets used in our experiments. We observe that most ratios are relatively low (less than 20%) across all datasets. This phenomenon is especially pronounced in datasets with a large number of labels, such as *bibtex*, and *COCO*, which is expected since instances are generally associated with only a few labels.

Additionally, we present the transformed multinomial distribution for the *VOC* dataset in Figure 2. The results indicate that most transformed categories have very few or even zero samples, posing a significant challenge for direct multinomial estimation.

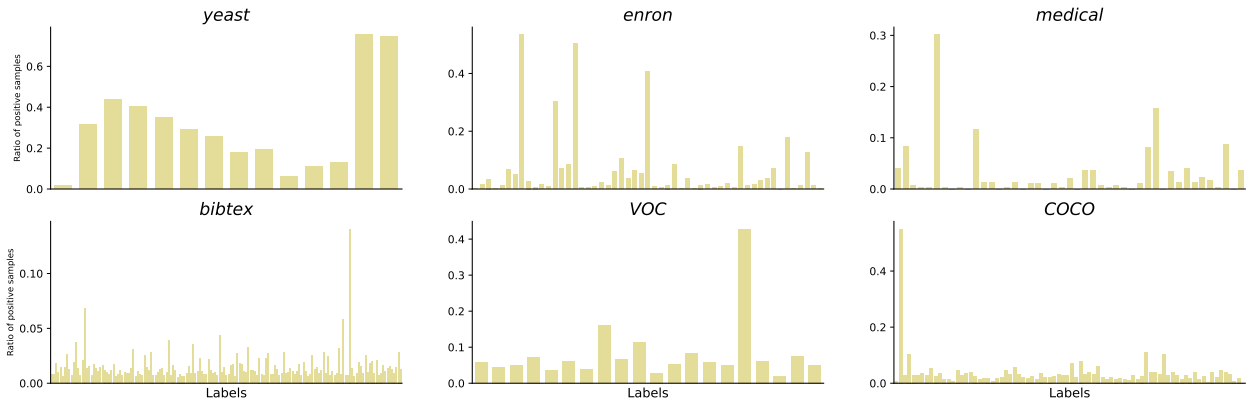

Figure 1. The ratio of positive samples for original label in different datasets.

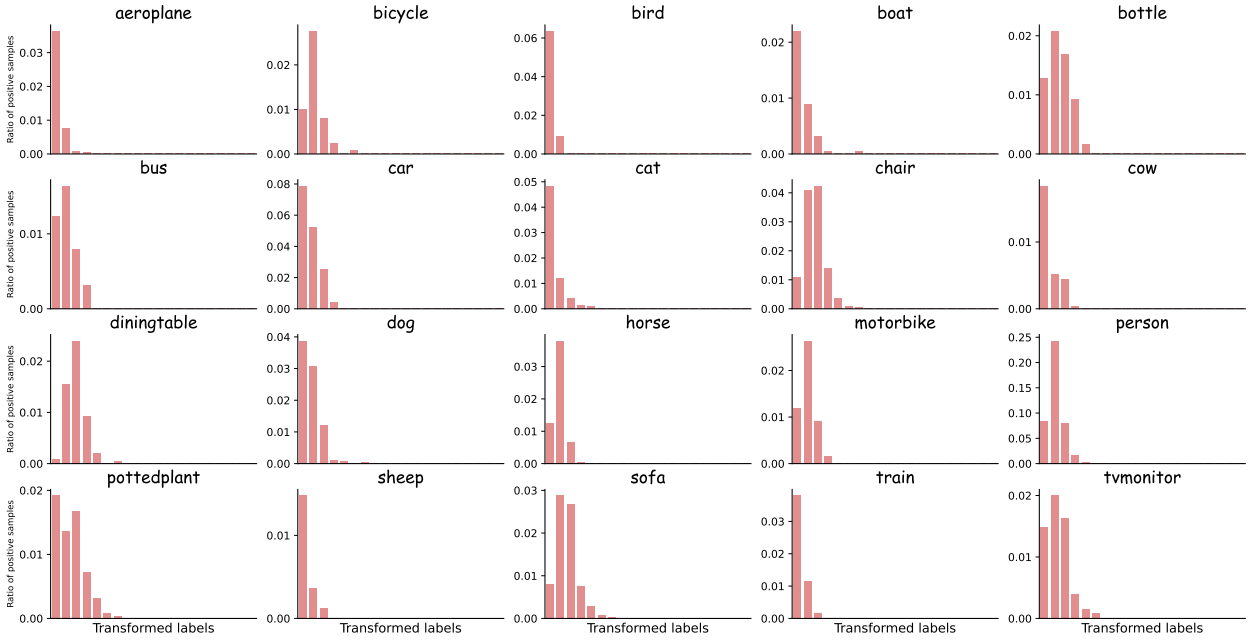

Figure 2. The transformed multinomial distribution in the *VOC* dataset. Each subplot corresponds to data with $Y_j = 1$ and represents the distribution of $\|\mathbf{Y}\|_1 = k$ for $k = 0, 1, \cdots, q$.

C. Implementation Details

Dataset Splitting. We adhere to the train-test splits provided by the *scikit-multilearn* package [1] for the *yeast*, *enron*, *medical*, and *bibtex* datasets. Additionally, we randomly reserve 10% of the training set as a validation set for hyperparameter tuning and early stopping. For the *VOC* dataset, we use its default train-val-test split. For the *COCO* dataset, we utilize its default validation set as the test set and randomly sample 10% of the training set as the validation set.

Feature Extraction. In order to save the computational resources, we employ the pre-trained ResNet-50 model from the *torchvision* [2] package to extract features for the *VOC* and *COCO* datasets. The final classification layer is removed, and the output of the global average pooling layer is used as the feature representation, yielding a 1000-dimensional feature vector for each image.

[1] http://scikit.ml/index.html
[2] https://pytorch.org/vision/stable/index.html

Table 1. Number of samples m used to estimate $\widehat{\mathbf{V}}$ for different datasets.

yeast	enron	medical	bibtex	VOC	COCO
200	100	200	100	150	150

Model Training We employ a feedforward neural network with two hidden layers, followed by q output heads. The sizes of the hidden layers are selected from $[256, 256]$ for *yeast* and $[512, 512]$ for the others. The Adam optimizer is used with a learning rate of 0.001, and the batch size is set to 1024. The model is trained for 200 epochs. The number of samples m used to estimate $\widehat{\mathbf{V}}$ is specified in Tab. 1. A dropout rate of 0.3 is applied across all datasets. The loss function is the binary cross-entropy loss. Early stopping is applied with a patience of 20 epochs on the validation set, monitored by the validation loss, to prevent overfitting.

D. Empirical Convergence Analysis

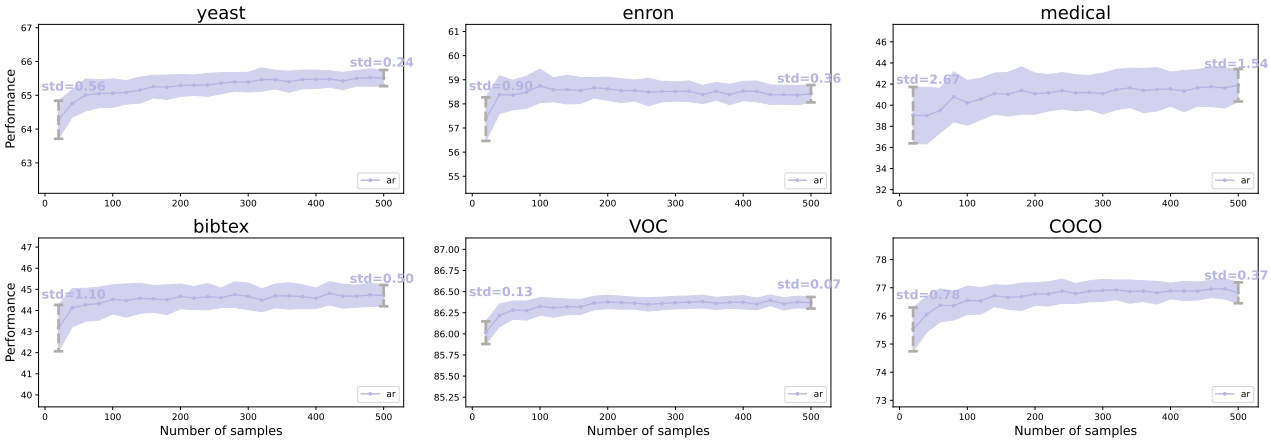

Figure 3. Empirical convergence analysis: the mean and std of performance by varying numbers of samples m.

Theorem 2 theoretically demonstrates that the estimation error of $\widehat{\mathbf{V}}$ decreases as the sample size m increases. In this section, we further investigate empirical convergence to intuitively evaluate the efficiency of the proposed method. Specifically, we vary the number of samples m from 20 to 500 with a step size 20, and plot the mean and std across 50 runs on 100 instances of the validation set, as shown in Figure 3.

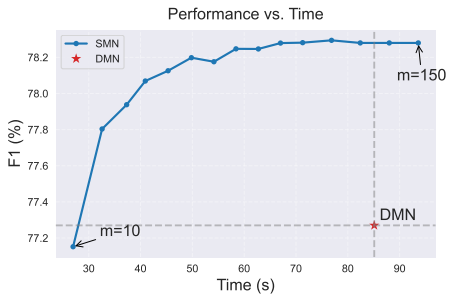

Figure 4. Runtime comparison by varying m on *COCO*.

A key observation is that performance converges within 150 samples for most datasets, with std of less than 0.5%. An exception is the *medical* dataset, which exhibits higher variance for reasons previously discussed. Nevertheless, this rapid convergence is observed even for datasets with a large number of labels, such as *bibtex* and *COCO*, likely due to the intrinsically low-dimensional structure of the label space and the label dependencies captured by the autoregressive model, both of which facilitate convergence.

To examine the runtime of SMN more closely as the sample size m varies, Figure 4 shows that, beyond a fixed initial overhead, the runtime increases steadily with m. More importantly, when anchoring to either same performance or same time, SMN provies a better accuracy-time trade-off than DMN. Note that DMN incurs higher overhead because its multinomial estimators are more expensive than the binary ones in SMN.

E. Additional results to "One model for many metrics"

In Section 6.6, we demonstrate that SMN achieves competitive performance across three different metrics using a unified model, by applying different inference procedures to the same set of samples. In Tab. 2, we present additional results on *VOC* and *COCO* to further support this finding. Note that Label Powerset (LP) requires estimating 2^q label sets, which becomes worse for *VOC* ($q = 20$) and not applicable (NA) for *COCO* ($q = 80$).

Table 2. Optimization for different metrics in *VOC* and *COCO*.

	DMN	BR	LP	Ours	DMN	BR	LP	Ours
	VOC				*COCO*			
F1	<u>85.01</u>	84.24	83.06	**86.08**	<u>77.27</u>	77.19	NA	**78.22**
HA	97.29	<u>97.67</u>	97.18	**97.69**	98.07	**98.35**	NA	<u>98.34</u>
SA	62.86	<u>65.00</u>	63.03	**65.39**	35.82	<u>37.48</u>	NA	**37.56**

References

[1] Krzysztof Dembczynski, Arkadiusz Jachnik, Wojciech Kotlowski, Willem Waegeman, and Eyke Hüllermeier. Optimizing the f-measure in multi-label classification: Plug-in rule approach versus structured loss minimization. In *International conference on machine learning*, pages 1130–1138. PMLR, 2013. 1