Bayes optimal prediction for NDCG@k in extreme multi-label classification

Kalina Jasinska[1] and **Krzysztof Dembczyński**[2]

Abstract. Extreme multi-label classification (XMLC) is a supervised learning problem in which instances are labeled with a few relevant labels from a very large set, consisting of potentially millions of target labels. For many performance metrics, the Bayes optimal prediction is a simple function of marginal label probabilities. This is a case, for example, of Hamming loss, macro F-measure, and precision@k. In this paper we analyze NDCG@k and show that under specific conditions its Bayes optimal prediction is also based on marginal probabilities. The open problem remains, however, how inferior is the decision rule based on marginal probabilities to the optimal prediction in the general case.

1 Problem statement

Let $\mathcal{X}$ denote a feature space and $\mathcal{L} = \{1, \ldots, m\}$ be a finite set of m class labels. Each subset of labels can be represented by a binary vector $\boldsymbol{y}$ of length m, with $y_i = 1$ if and only if label i is in the subset. We use $|\boldsymbol{y}|$ to denote the number of labels in the vector $\boldsymbol{y}$, i.e., $\sum_i y_i$. The label vector space is denoted by $\mathcal{Y}$. Observations $(\boldsymbol{x}, \boldsymbol{y})$ are generated independently and identically according to the probability distribution $\mathbf{P}(\boldsymbol{X} = \boldsymbol{x}, \boldsymbol{Y} = \boldsymbol{y})$ (denoted later by $\mathbf{P}(\boldsymbol{x}, \boldsymbol{y})$) defined on $\mathcal{X} \times \mathcal{Y}$.

Formally, the XMLC problem can be defined as finding a *classifier* $\boldsymbol{h}(\boldsymbol{x}) = (h_1(\boldsymbol{x}), \ldots, h_m(\boldsymbol{x}))$, defined as a mapping $\mathcal{X} \to \mathcal{R}^m$, that minimizes the *expected loss*:

$$L_\ell(\boldsymbol{h}) = \mathbb{E}_{(\boldsymbol{x},\boldsymbol{y})\sim\mathbf{P}(\boldsymbol{x},\boldsymbol{y})}(\ell(\boldsymbol{y}, \boldsymbol{h}(\boldsymbol{x}))\,,$$

where $\ell(\boldsymbol{y}, \hat{\boldsymbol{y}})$ is the (*task*) *loss*. The expected loss for a single $\boldsymbol{x}$ $\mathbb{E}\ell_{\log}(\boldsymbol{h} \,|\, \boldsymbol{x})$ is called a *conditional risk*. The optimal classifier, the so-called *Bayes classifier*, for a given loss function ℓ is:

$$\boldsymbol{h}^*_\ell \in \arg\min_{\boldsymbol{h}} L_\ell(\boldsymbol{h})\,.$$

2 Performance measures

In the following we focus on NDCG@k, but before discussing this performance measure, let us first recall two other related measures, precision@k and DCG@k. All considered measures are utilities, so to use them in the formal framework defined above we will changed them to loss functions, whenever it is necessary.

Precision@k is defined as:

$$p@k(\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{h}) = \frac{1}{k} \sum_{j \in \rho_k(\boldsymbol{h})} [\![y_j = 1]\!]\,, \tag{1}$$

where $\rho_k(\boldsymbol{h})$ is a set of k labels predicted by $\boldsymbol{h}$ for $\boldsymbol{x}$. Usually, the labels can be predicted as top k labels from some returned ranking, but it is worth to notice that the definition of precision@k does not require the labels to be sorted. This is different for DCG@k and NDCG@k. Therefore, in this case we define permutation σ of labels returned by classifier $\boldsymbol{h}$ such that $\sigma(\boldsymbol{h}, r) \in \mathcal{L}$ is a label on the r-th rank. By extending the definition of precision@k with a discounting factor $\frac{1}{\log_2(r+1)}$ and removing $\frac{1}{k}$ we get the discounted cumulative gain:

$$DCG@k(\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{h}) = \sum_{r=1}^{k} \frac{y_{\sigma(\boldsymbol{h},r)}}{\log_2(r+1)}. \tag{2}$$

The best possible, or the ideal, DCG@k for a given label vector $\boldsymbol{y}$ is

$$IDCG@k(\boldsymbol{y}) = \sum_{r=1}^{\min(k,|\boldsymbol{y}|)} \frac{1}{\log_2(r+1)},$$

where $|\boldsymbol{y}|$ denotes the number of positive elements in $\boldsymbol{y}$. By normalizing DCG@k by this factor we get the normalized discounted cumulative gain

$$NDCG@k(\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{h}) = N_k(\boldsymbol{y}) DCG@k(\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{h}), \tag{3}$$

where, for simplicity of notation, we define $N_k(\boldsymbol{y}) = IDCG^{-1}(\boldsymbol{y})$.

Many multi-label learning algorithms are suited for delivering estimates of marginal probabilities

$$\eta_j(\boldsymbol{x}) = \mathbf{P}(y_j = 1 \,|\, \boldsymbol{x}) = \sum_{\boldsymbol{y}: y_j = 1} \mathbf{P}(\boldsymbol{y}, \boldsymbol{x})\,.$$

Sorting the labels by these estimates and selecting top values is an easy way of obtaining final predictions. In this paper we analyze the consistency of such approach for aforementioned metrics.

3 Bayes optimal predictions

In this section, we show that the Bayes optimal predictions for precision@k and DCG@k are indeed based on marginal probabilities. This is, however, not the case for NDCG@k.

Let us start by recalling a result regarding precision@k from [5].

Theorem 1. *Predicting k labels with the highest marginal probabilities $\eta_j(\boldsymbol{x})$ gives an optimal solution for precision@k.*

Proof. Let us first define a precision@k-based loss as

$$\ell_{p@k}(\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{h}) = -p@k\,.$$

[1] Poznan University of Technology, email: kjasinska@cs.put.poznan.pl
[2] Poznan University of Technology, email: kdembczynski@cs.put.poznan.pl

The conditional risk can be then given as:

$$\begin{aligned} L_{p@k}(\boldsymbol{h} \mid \boldsymbol{x}) &= \mathbb{E}\ell_{p@k}(\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{h}) \\ &= -\sum_{\boldsymbol{y}\in\mathcal{Y}} \mathbf{P}(\boldsymbol{y} \mid \boldsymbol{x}) \frac{1}{k} \sum_{j\in\hat{\mathcal{Y}}_k} [\![y_j = 1]\!] \\ &= -\frac{1}{k} \sum_{j\in\hat{\mathcal{Y}}_k} \sum_{\boldsymbol{y}\in\mathcal{Y}} \mathbf{P}(\boldsymbol{y} \mid \boldsymbol{x}) [\![y_j = 1]\!] \\ &= -\frac{1}{k} \sum_{j\in\hat{\mathcal{Y}}_k} \eta_j(\boldsymbol{x}) \,. \end{aligned}$$

This value is minimized when $\sum_{j\in\hat{\mathcal{Y}}_k} \eta_j(\boldsymbol{x})$ is maximized. The maximum sum is obtained by choosing k labels with the highest marginal probabilities $\eta_j(\boldsymbol{x})$. □

As it turns out, the optimal strategy for DCG@k is similar to the one for precision@k, however, it requires additionally sorting by marginal probabilities to obtain the optimal order.

Theorem 2. *Predicting k labels with the highest marginal probabilities $\eta_j(\boldsymbol{x})$ sorted in descending order by the marginal probability gives an optimal solution for DCG@k.*

The proof is analogous to the one for precision@k. It is given in the Appendix A. Let us denote a classifier that delivers such answer with $\boldsymbol{h}^*_{D@k}$. Notice that this classifier is also optimal for precision@k. Interestingly, for $k = m$, it also coincides with the optimal solution for the unormalized rank loss [1].

For NDCG@k the optimal strategy is different, however also consists of sorting the labels according to specific marginalized values.

Theorem 3. *Let $\Delta_j(k, \boldsymbol{x})$ be the following marginalized value:*

$$\Delta_j(k, \boldsymbol{x}) = \sum_{\boldsymbol{y}: y_j=1} N_k(\boldsymbol{y}) \mathbf{P}(\boldsymbol{y}|\boldsymbol{x}). \tag{4}$$

Predicting k labels with the highest values of $\Delta_j(k, \boldsymbol{x})$ sorted in descending order by this value gives an optimal solution for NDCG@k.

The proof, analogous to the previous ones, is given in the Appendix B. Let us denote a classifier that delivers such answer with $\boldsymbol{h}^*_{N@k}$. Notice two important properties of NDCG@k. Firstly, the optimal decision is based on specific marginal quantities that take the number of labels into account. This is similar to the Bayes decisions for normalized rank loss [2] and instance-wise F-measure [4]. Notice also that for NDCG@k, label vectors with smaller $|\boldsymbol{y}|$ contribute more to $\Delta_j(k, \boldsymbol{x})$.

Secondly, the optimal decision depends on k, i.e., for a fixed conditional distribution, NDCG@k and NDCG@l for $k < l$ can be optimized by different rankings on top k positions. To be more precise, it does not necessarily hold that $\forall_{r=1,\ldots,k} \boldsymbol{h}^*_{N@k}(r) = \boldsymbol{h}^*_{N@l}(r)$, where $\boldsymbol{h}_{N@k}(r)$ denotes the r-th prediction according to decreasing $\Delta_j(k, \boldsymbol{x})$. This issue is described in the Appendix C.

4 Approximating $\boldsymbol{h}^*_{N@k}$ by $\boldsymbol{h}^*_{D@k}$

As we have seen, $\boldsymbol{h}^*_{D@k}$ and $\boldsymbol{h}^*_{N@k}$ are in general different solutions. In this section, we try to answer a question how good is $\boldsymbol{h}^*_{D@k}$ in approximating $\boldsymbol{h}^*_{N@k}$. This is of particular interest as $\boldsymbol{h}^*_{D@k}$ can be much easier estimated as it is based on sorting marginal probabilities. The exhaustive answer for this question is not simple, but we can easily indicate two specific situations in which $\boldsymbol{h}^*_{D@k} = \boldsymbol{h}^*_{N@k}$.

4.1 Conditionally independent labels

Let us first consider the case of conditionally independent labels. The conditional joint probability distribution is then expressed by:

$$\mathbf{P}(\boldsymbol{y}|\boldsymbol{x}) = \prod_{i=1}^{m} \eta_i(\boldsymbol{x})^{y_i} (1 - \eta_i(\boldsymbol{x}))^{1-y_i}.$$

Theorem 4. *Given conditionally independent labels, a classifier $\boldsymbol{h}^*_{D@k}$ delivers optimal solution for NDCG@k.*

A detailed proof is given in the Appending D.

4.2 NDCG@1

The conditional independence assumption may not necessarily hold. However, it is easy to show that for any distribution, the optimal solution for NDCG@1 is to select the label with the highest marginal probability.

Theorem 5. *A classifier $\boldsymbol{h}^*_{D@1}$ delivers an optimal solution for NDCG@1.*

Proof. To proof this theorem we will show that $\Delta_j(1, \boldsymbol{x}) = \eta_j(\boldsymbol{x})$, for $j = 1, \ldots, m$. Consider equation (4):

$$\Delta_j(1, \boldsymbol{x}) = \sum_{\boldsymbol{y}: y_j=1} N_1(\boldsymbol{y}) \mathbf{P}(\boldsymbol{y}|\boldsymbol{x}),$$

where $N_1(\boldsymbol{y}) = \left(\sum_{r=1}^{\min(1,|\boldsymbol{y}|)} \frac{1}{\log_2(r+1)} \right)^{-1}$.

Consider two cases. If $|\boldsymbol{y}| = 0$, then this element does not contribute to $\Delta_j(1, \boldsymbol{x})$, since $y_j = 0$ (also $N_1(\boldsymbol{y})$ equals 0 in this case). If $|\boldsymbol{y}| \geq 1$, then $N_1(\boldsymbol{y})$ is always equal to $\frac{1}{\log_2(2)} = 1$. Therefore

$$\Delta_j(1, \boldsymbol{x}) = \sum_{\boldsymbol{y}: y_j=1} N_1(\boldsymbol{y}) \mathbf{P}(\boldsymbol{y}|\boldsymbol{x}) = \sum_{\boldsymbol{y}: y_j=1} \mathbf{P}(\boldsymbol{y}|\boldsymbol{x}) = \eta_j(\boldsymbol{x}).$$

□

4.3 General case

For the future work we plan to analize the link between $\boldsymbol{h}^*_{D@k}$ and $\boldsymbol{h}^*_{N@k}$ in the general case. In other words, we would like to bound the difference $L_{N@k}(\boldsymbol{h}^*_{D@k}|\boldsymbol{x}) - L_{N@k}(\boldsymbol{h}^*_{N@k}|\boldsymbol{x})$ for an arbitrary k and distribution. We will start, however, with the results for small values of k. This analysis will complete the NDCG consistency results for listwise rankings obtained in [3].

Acknowledgments

This work was supported by the Polish National Science Center under grant no. 2017/25/N/ST6/00747.

REFERENCES

[1] K. Dembczyński, W. Cheng, and E. Hüllermeier, 'Bayes optimal multilabel classification via probabilistic classifier chains', in *ICML*, pp. 279–286, (2010).

[2] K. Dembczyński, W. Kotłowski, and E. Hüllermeier, 'Consistent multilabel ranking through univariate losses', in *ICML*, pp. 1347–1354, (2012).

[3] P. Ravikumar, A. Tewari, and E. Yang, 'On NDCG consistency of listwise ranking methods', in *AISTATS*, pp. 618–626, (2011).

[4] W. Waegeman, K. Dembczynski, A. Jachnik, W. Cheng, and E. Hüllermeier, 'On the Bayes-optimality of F-measure maximizers', *Journal of Machine Learning Research*, **15**(1), 3333–3388, (2014).

[5] M. Wydmuch, K. Jasinska, K. Dembczyński, M. Kuznetsov, and R. Busa-Fekete, 'A no-regret generalization of hierarchical softmax to extreme multi-label classification', in *NIPS*, (2018).

A DCG@k

Theorem 2. *Predicting k labels with the highest marginal probabilities $\eta_j(\boldsymbol{x})$ sorted in descending order by the marginal probability gives an optimal solution for DCG@k.*

Proof. The proof is analogous to the one for precision@k. Let $\ell_{D@k} = -\mathrm{DCG@}k(\boldsymbol{y}, \boldsymbol{z}, \boldsymbol{h})$ denote the loss function corresponding to DCG@k. The conditional risk is then

$$\begin{aligned} L_{D@k}(\boldsymbol{h} \,|\, \boldsymbol{x}) &= \mathbb{E}\ell_{D@k}(\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{h}) \\ &= -\sum_{\boldsymbol{y}\in\mathcal{Y}} \mathbf{P}(\boldsymbol{y}|\boldsymbol{x}) \sum_{r=1}^{k} \frac{y_{\sigma(\boldsymbol{h},r)}}{\log_2(r+1)} \\ &= -\sum_{r=1}^{k} \frac{1}{\log_2(r+1)} \sum_{\boldsymbol{y}\in\mathcal{Y}} y_{\sigma(\boldsymbol{h},r)} \mathbf{P}(\boldsymbol{y}|\boldsymbol{x}) \end{aligned}$$

This value is minimized when

$$\sum_{r=1}^{k} \frac{1}{\log_2(r+1)} \sum_{\boldsymbol{y}\in\mathcal{Y}} y_{\sigma(\boldsymbol{h},r)} \mathbf{P}(\boldsymbol{y}|\boldsymbol{x})$$

is maximized. Notice that because $y_{\sigma(\boldsymbol{h},r)} \in \{0, 1\}$, we have $\sum_{\boldsymbol{y}\in\mathcal{Y}} y_{\sigma(\boldsymbol{h},r)} \mathbf{P}(\boldsymbol{y}|\boldsymbol{x}) = \eta_j(\boldsymbol{x})$. Moreover, the discounted gains $\frac{1}{\log_2(r+1)}$ diminish with growing r. Therefore to maximize DCG@k one should rank labels by $\eta_j(\boldsymbol{x})$. □

B NDCG@k

Theorem 3. *Let $\Delta_j(k, \boldsymbol{x})$ be the following marginalized value:*

$$\Delta_j(k, \boldsymbol{x}) = \sum_{\boldsymbol{y}: y_j = 1} N_k(\boldsymbol{y}) \mathbf{P}(\boldsymbol{y}|\boldsymbol{x}). \tag{4}$$

Predicting k labels with the highest values of $\Delta_j(k, \boldsymbol{x})$ sorted in descending order by this value gives an optimal solution for NDCG@k.

Proof. Let $\ell_{N@k}(\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{h}) = -\mathrm{NDCG@k}(\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{h})$ be a loss function corresponding to NDCG@k. Then, the conditional risk is:

$$\begin{aligned} L_{N@k}(\boldsymbol{h} \,|\, \boldsymbol{x}) &= \mathbb{E}\ell_{N@k}(\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{h}) \\ &= -\sum_{\boldsymbol{y}\in\mathcal{Y}} \mathbf{P}(\boldsymbol{y}|\boldsymbol{x}) N_k(\boldsymbol{y}) \sum_{r=1}^{k} \frac{y_{\sigma(\boldsymbol{h},r)}}{\log_2(r+1)} \\ &= -\sum_{r=1}^{k} \frac{1}{\log_2(r+1)} \sum_{\boldsymbol{y}\in\mathcal{Y}} y_{\sigma(\boldsymbol{h},r)} \mathbf{P}(\boldsymbol{y}|\boldsymbol{x}) N_k(\boldsymbol{y}) \\ &= -\sum_{r=1}^{k} \frac{1}{\log_2(r+1)} \Delta_j(k, \boldsymbol{x}) \end{aligned}$$

Analogously to Theorem 2, we see that the risk is minimized by top k labels sorted decreasingly by the values of $\Delta_j(k, \boldsymbol{x})$. □

C Inconsistency of rankings for NDCG@k and NDCG@l with $k < l$

Theorem 6. *Let $k < l \leq m$. It does not necessarily hold that $\forall_{r=1,\ldots,k} \boldsymbol{h}^*_{N@k}(r) = \boldsymbol{h}^*_{N@l}(r)$, where $\boldsymbol{h}_{N@k}(r)$ denotes the r-th prediction according to decreasing $\Delta_j(k, \boldsymbol{x})$.*

Proof. This can be proved by a counterexample for $k = 1$ and $l = 2$. Let the joint distribution be non-zero for $\boldsymbol{y}_1 = (1, 0, 0)$ and $\boldsymbol{y}_2 = (0, 1, 1)$, with marginal probabilities $\eta_1(\boldsymbol{x}) < \eta_2(\boldsymbol{x}) = \eta_3(\boldsymbol{x})$. Then, the optimal decision for NDCG@1 is a ranking consisting of one label, being either y_2 or y_3, as their marginal probabilities are the same. A joint probability for which the ranking on the first position is different for NDCG@2 has to satisfy:

$$\Delta_1(2, \boldsymbol{x}) = N_2(\boldsymbol{y}_1)\eta_1(\boldsymbol{x}) = \eta_1(\boldsymbol{x}) > \Delta_2(2, \boldsymbol{x}) = N_2(\boldsymbol{y}_2)\eta_2(\boldsymbol{x}).$$

In the above $N_2(\boldsymbol{y}_1) = 1$, since $|\boldsymbol{y}_1| = 1$. It easy to see, for example, that $\eta_1(\boldsymbol{x}) = 0.4$ and $\eta_2(\boldsymbol{x}) = 0.6$ fulfill the requirements as $N_2(\boldsymbol{y}_2) = \left(\frac{1}{\log_2(2)} + \frac{1}{\log_2(3)}\right)^{-1} = \frac{\log_2(3)}{1+\log_2(3)}$ and $N_2(\boldsymbol{y}_2) \cdot 0.6 \approx 0.3679 < 0.4$. Therefore the first element of the ranking optimal for NDCG@2 is y_1, which is different from the optimal decision for NDCG@1. □

Notice that it is impossible to construct such example for $m = 2$. It is so because either $\boldsymbol{y}$ has length 1 (so its probability contributes to marginal probabilities as strong as to Δ) or $\boldsymbol{y}$ has length 2 (and then the value contributes to marginal values for those two labels equally).

D NDCG@k and conditionally independent labels

Theorem 4. *Given conditionally independent labels, a classifier $\boldsymbol{h}^*_{D@k}$ delivers optimal solution for NDCG@k.*

Proof. To proof the theorem it suffices to show that for conditionally independent labels the order of labels induced by the marginal probabilities $\eta_j(\boldsymbol{x})$ is the same as the order induced by the values of $\Delta_j(k, \boldsymbol{x})$. In other words, for any two labels $i, j \in \{1, \ldots, m\}$, $i \neq j$, and for any $k \in \mathcal{L}$, $\eta_i(\boldsymbol{x}) \geq \eta_j(\boldsymbol{x}) \Leftrightarrow \Delta_i(k, \boldsymbol{x}) \geq \Delta_j(k, \boldsymbol{x})$.

Let $\eta_i(\boldsymbol{x}) \geq \eta_j(\boldsymbol{x})$. Then in the summation over all $\boldsymbol{y}$ in (4), which can be rewritten in the following way

$$\Delta_j(k, \boldsymbol{x}) = \sum_{\boldsymbol{y} \in \mathcal{Y}} y_j N_k(\boldsymbol{y}) \mathbf{P}(\boldsymbol{y}|\boldsymbol{x}),$$

we consider four subsets of $\mathcal{Y}$, creating a partition of this set:

$$\mathcal{S}^{u,w}_{i,j} = \{\boldsymbol{y} \in \mathcal{Y} : y_i = u \wedge y_j = w\}, u, w \in \{0, 1\}.$$

The subset $\mathcal{S}^{0,0}_{i,j}$ does not play any role because $y_i = y_j = 0$ and therefore do not contribute to the final sum. Then (4) can be written in the following way:

$$\Delta_i(k, \boldsymbol{x}) = \sum_{\boldsymbol{y}: \mathcal{S}^{1,0}_{i,j}} N_k(\boldsymbol{y}) \mathbf{P}(\boldsymbol{y}|\boldsymbol{x}) + \sum_{\boldsymbol{y} \in \mathcal{S}^{1,1}_{i,j}} N_k(\boldsymbol{y}) \mathbf{P}(\boldsymbol{y}|\boldsymbol{x}) \tag{5}$$

$$\Delta_j(k, \boldsymbol{x}) = \sum_{\boldsymbol{y}: \mathcal{S}^{0,1}_{i,j}} N_k(\boldsymbol{y}) \mathbf{P}(\boldsymbol{y}|\boldsymbol{x}) + \sum_{\boldsymbol{y} \in \mathcal{S}^{1,1}_{i,j}} N_k(\boldsymbol{y}) \mathbf{P}(\boldsymbol{y}|\boldsymbol{x}) \tag{6}$$

The contribution of elements from $\mathcal{S}^{1,1}_{i,j}$ is equal for both $\Delta_i(k, \boldsymbol{x})$ and $\Delta_j(k, \boldsymbol{x})$. It is so because the value of $N_k(\boldsymbol{y})\mathbf{P}(\boldsymbol{y}|\boldsymbol{x})$ is the same for all $\boldsymbol{y} \in \mathcal{S}^{1,1}_{i,j}$: the conditional joint probabilities $\mathbf{P}(\boldsymbol{y}|\boldsymbol{x})$ are fixed and they are multiplied by the same factors $N_k(\boldsymbol{y})$.

Consider now the contributions of $\mathcal{S}^{1,0}_{i,j}$ and $\mathcal{S}^{0,1}_{i,j}$ to the relevant sums. By the definition of $\mathcal{Y}$, $\mathcal{S}^{1,0}_{i,j}$, and $\mathcal{S}^{0,1}_{i,j}$, there exists bijection $b_{i,j} : \mathcal{S}^{1,0}_{i,j} \to \mathcal{S}^{0,1}_{i,j}$, such that for each $\boldsymbol{y}' \in \mathcal{S}^{1,0}_{i,j}$ there exists $\boldsymbol{y}'' \in \mathcal{S}^{0,1}_{i,j}$ equal to $\boldsymbol{y}'$ except on the i-th and the j-th position.

Notice that because of the conditional independence assumption the joint probabilities of elements in $\mathcal{S}^{1,0}_{i,j}$ and $\mathcal{S}^{0,1}_{i,j}$ are related to each other. Let $\boldsymbol{y}'' = b_{i,j}(\boldsymbol{y}')$, where $\boldsymbol{y}' \in \mathcal{S}^{1,0}_{i,j}$ and $\boldsymbol{y}'' \in \mathcal{S}^{0,1}_{i,j}$. The joint probabilities are:

$$\mathbf{P}(\boldsymbol{y}'|\boldsymbol{x}) = \eta_i(\boldsymbol{x})(1 - \eta_j(\boldsymbol{x})) \prod_{l \in \mathcal{L} \setminus \{i,j\}} \eta_l(\boldsymbol{x})^{y_l} (1 - \eta_l(\boldsymbol{x}))^{1 - y_l}$$

and

$$\mathbf{P}(\boldsymbol{y}''|\boldsymbol{x}) = (1 - \eta_i(\boldsymbol{x}))\eta_j(\boldsymbol{x}) \prod_{l \in \mathcal{L} \setminus \{i,j\}} \eta_l(\boldsymbol{x})^{y_l} (1 - \eta_l(\boldsymbol{x}))^{1 - y_l}.$$

One can easily notice the relation between these probabilities:

$$\mathbf{P}(\boldsymbol{y}'|\boldsymbol{x}) = \eta_i(\boldsymbol{x})(1 - \eta_j(\boldsymbol{x}))q_{i,j}$$

and

$$\mathbf{P}(\boldsymbol{y}''|\boldsymbol{x}) = (1 - \eta_i(\boldsymbol{x}))\eta_j(\boldsymbol{x})q_{i,j},$$

where $q_{i,j} = \prod_{l \in \mathcal{L} \setminus \{i,j\}} \eta_l(\boldsymbol{x})^{y_l} (1 - \eta_l(\boldsymbol{x}))^{1 - y_l} \geq 0$. Consider now the difference of these two probabilities:

$$\begin{aligned}
\mathbf{P}(\boldsymbol{y}'|\boldsymbol{x}) - \mathbf{P}(\boldsymbol{y}''|\boldsymbol{x}) &= \eta_i(\boldsymbol{x})(1 - \eta_j(\boldsymbol{x}))q_{i,j} - (1 - \eta_i(\boldsymbol{x}))\eta_j(\boldsymbol{x})q_{i,j} \\
&= q_{i,j}(\eta_i(\boldsymbol{x})(1 - \eta_j(\boldsymbol{x})) - (1 - \eta_i(\boldsymbol{x}))\eta_j(\boldsymbol{x})) \\
&= q_{i,j}(\eta_i(\boldsymbol{x}) - \eta_i(\boldsymbol{x})\eta_j(\boldsymbol{x}) - \eta_j(\boldsymbol{x}) + \eta_i(\boldsymbol{x})\eta_j(\boldsymbol{x})) \\
&= q_{i,j}(\eta_i(\boldsymbol{x}) - \eta_j(\boldsymbol{x})).
\end{aligned}$$

From the above we see that $\eta_i(\boldsymbol{x}) \geq \eta_j(\boldsymbol{x}) \Rightarrow \mathbf{P}(\boldsymbol{y}'|\boldsymbol{x}) \geq \mathbf{P}(\boldsymbol{y}''|\boldsymbol{x})$. Due to the properties of the bijection $b_{i,j}$, the number of positive labels in $\boldsymbol{y}'$ and $\boldsymbol{y}''$ is the same and $N_k(\boldsymbol{y}') = N_k(\boldsymbol{y}'')$, therefore we also get $\eta_i(\boldsymbol{x}) \geq \eta_j(\boldsymbol{x}) \Rightarrow \sum_{\boldsymbol{y}: \mathcal{S}^{1,0}_{i,j}} N_k(\boldsymbol{y})\mathbf{P}(\boldsymbol{y}|\boldsymbol{x}) \geq \sum_{\boldsymbol{y}: \mathcal{S}^{0,1}_{i,j}} N_k(\boldsymbol{y})\mathbf{P}(\boldsymbol{y}|\boldsymbol{x})$, which finally based on (5) and (6) gives us $\eta_i(\boldsymbol{x}) \geq \eta_j(\boldsymbol{x}) \Rightarrow \Delta_i(k, \boldsymbol{x}) \geq \Delta_j(k, \boldsymbol{x})$.

The implication in the other side, i.e., $\eta_i(\boldsymbol{x}) \geq \eta_j(\boldsymbol{x}) \Leftarrow \mathbf{P}(\boldsymbol{y}'|\boldsymbol{x}) \geq \mathbf{P}(\boldsymbol{y}''|\boldsymbol{x})$ holds obviously for $q_{i,j} > 0$. For $q_{i,j} = 0$, we can notice, however, that $\mathbf{P}(\boldsymbol{y}'|\boldsymbol{x})$ and $\mathbf{P}(\boldsymbol{y}''|\boldsymbol{x})$ do not contribute to the appropriate sums as they are zero, and therefore we can follow a similar reasoning as above, concluding that $\eta_i(\boldsymbol{x}) \geq \eta_j(\boldsymbol{x}) \Leftarrow \Delta_i(k, \boldsymbol{x}) \geq \Delta_j(k, \boldsymbol{x})$.

Thus for conditionally independent labels, the order of labels induced by marginal probabilities $\eta_j(\boldsymbol{x})$ is equal to the order induced by $\Delta_j(k, \boldsymbol{x})$. By the optimality of the order induced by $\Delta_j(k, \boldsymbol{x})$, proven in Theorem 3, for conditionally independent labels the order induced by the marginal probabilities is optimal for NDCG@k. □