Supplementary Material for Conformal Prediction Sets for Ordinal Classification

APPENDIX

Errata

Below are a list of important corrections that we discovered while reviewing the submitted version of our paper. The proofs in Appendix B include the corrected statements of the theorems.

- **Section 4 - Line 125:** $\hat{q}_{D_{cal}}(\alpha)$ is the score threshold defined as the bias-adjusted $(\alpha)^{th}$ quantile of the model score of the true label and not $(1-\alpha)^{th}$.
- **Theorem 1:** The claim $|\hat{S}_{D,\alpha}(\mathbf{x})| \leq |S^{oracle}_{\alpha-2\delta}(\mathbf{x})|$ should be replaced by $|\hat{S}_{D,\alpha}(\mathbf{x})| \leq |S^{oracle}_{\alpha-4\delta-\frac{1}{n+1}}(\mathbf{x})|$ where n is the size of the calibration set.
- **Theorem 2- Part (b):** The assumption on $\phi(\cdot)$ being surjective on R^+ was missed out. Further, the claim should be on the existence of a well-defined $\eta(\mathbf{x})$ and not uniqueness. $\eta(\mathbf{x})$ is unique only when $\phi(\cdot)$ is a bijective function such as $\phi(x) = \exp(x)$.

A Broader Impact

Our current work on contiguous conformal predictions for ordinal classification is foundational in nature and has multiple real-world applications.

- **Cancer Diagnosis.** Given the huge costs of misprediction for high-stakes applications such as cancer diagnosis, instead of a single point prediction it is useful to predict a contiguous set. For instance, prediction set of [`stage 2`, `stage 3`] gives a better notion of severity than a non-contiguous set such as [`no cancer`, `stage 3`] which might be discordant or a point prediction with low accuracy.
- **Dynamic Product Search Filters.** Customers new to any e-commerce platform often experience heavy cognitive load in specifying their requirements via search filters (e.g., budget, product dimensions). Identifying a small highly likely set of options based on their typical profile or immediate session history would significantly enhance the usability of the search filters and improve the customer experience.
- **Personalised Fit Recommendations.** Shopping apparel and shoes at any e-commerce platform is often tedious due to the limited support for fit-based recommendations. Often, customers find that the recommended products do not have options for their size and are forced to use search filters, which need to be repeatedly specified for each query. Additionally, customers also tend to order multiple products in the same size (bracketing) that results in a high return rate and excessive shipping costs for the platform. Automatic identification of the likely size ranges of a customer would improve the accuracy of recommendations and reduce shopping effort as well as return rates.
- **Personalised Budget Recommendations,** Since budget ranges have a natural ordering, automatic personalisation of product and brand recommendations for customers based on their preferred budget ranges is another area that can leverage COPOC to improve customer satisfaction.
- **Abuse Incident Audits.** E-commerce abuse incidents are often categorised along severity levels that have a natural ordering. Typically, human auditors are required to audit the abuse incidents, but current models do not often distinguish between a high chance of low severity incident vs. moderate chance of high severity incident. Conformal predictions can help streamline the audit workflows to better focus on the high severity incidents and optimise the overall outcomes both for e-commerce platform and the customers through expedited resolution.

B Theoretical Analysis

Lemma 1. *Given a fitted unimodal model $\hat{P}_{Y|X}$, for any test $\mathbf{x}$ and $\alpha \in (0,1]$, prediction sets $\hat{S}_{D,\alpha}(\mathbf{x})$ constructed using Eqn.6 or 7 has at least one solution which is contiguous (amongst multiple possibilities). When $\hat{P}_{Y|X}$ is strictly unimodal, all the solutions are contiguous.*

Proof: Let $\hat{S}_{D,\alpha}(\mathbf{x})$ be the prediction set with the shortest span (i.e., difference between highest and lowest included labels) as per Eqn.6 or 7.

Let $l+1$ and u denote the smallest and largest indices of the labels included in $\hat{S}_{D,\alpha}(\mathbf{x})$ so that the span is given by $u-l$.

Assuming $\hat{S}_{D,\alpha}(\mathbf{x})$ is non-contiguous implies that there exists at least one k^{skip} such that $(l+1) < k^{skip} < u$ and $c_{k^{skip}} \notin \hat{S}_{D,\alpha}(\mathbf{x})$. Let $\hat{p}_k(\mathbf{x}) = \hat{P}_{Y|X}(Y = c_k|X = \mathbf{x})$. Since $\hat{p}(\mathbf{x})$ is unimodal, there are two possible scenarios depending on where k^{skip} relies relative to the mode $c_{\hat{m}}$ of $\hat{p}$:

- $k^{skip} < \hat{m}$**:** In this case, we have $\hat{p}_{k^{skip}} \geq \hat{p}_{l+1}$ since $\hat{p}$ is non-decreasing before the mode
- $k^{skip} \geq \hat{m}$**:** In this case, we have $\hat{p}_{k^{skip}} \geq \hat{p}_u$ since $\hat{p}$ is non-increasing after the mode

Thus, $\hat{p}_{k^{skip}} \geq \min(\hat{p}_{l+1}, \hat{p}_u)$.

Case 1: LAC- PS Construction follows Eqn. 6: For this case, we have,

$$\hat{S}_{D,\alpha}(\mathbf{x}) = \{c_k \in \mathcal{C} | \hat{p}_k \geq \hat{q}_{D_{cal}}(\alpha)\}, \tag{6}$$

where $\hat{q}_{D_{cal}}(\alpha)$ is the bias-adjusted $(\alpha)^{th}$ quantile of the model score of the true label. Since c_{l+1} and c_u are included in $\hat{S}_{D,\alpha}(\mathbf{x})$, it follows that both $\hat{p}_{l+1} \geq \hat{q}_{D_{cal}}(\alpha)$ and $\hat{p}_u \geq \hat{q}_{D_{cal}}(\alpha)$.

Since $\hat{p}_{k^{skip}} \geq \min(\hat{p}_{l+1}, \hat{p}_u)$, it follows that $\hat{p}_{k^{skip}} \geq \hat{q}_{D_{cal}}(\alpha)$ as well implying that $c_{k^{skip}} \in \hat{S}_{D,\alpha}(\mathbf{x})$ which leads to a contradiction. Hence, the shortest span $\hat{S}_{D,\alpha}(\mathbf{x})$ has to be contiguous.

Case 2: APS- PS Construction follows Eqn. 7: For this case, we have,

$$\hat{S}_{\mathcal{D},\alpha}(\mathbf{x}) = \{c_{\pi_1}, c_{\pi_2} \ldots c_{\pi_j}\} \text{ where } j = sup \left\{ j\prime : \sum_{k=1}^{j\prime} \hat{p}_{\pi_k} < \hat{q}_{D_{cal}}(\alpha) \right\} + 1, \tag{7}$$

where π is a permutation of $\{1, \ldots, K\}$ that sorts $\hat{p}_k$ in the descending order from most likely to least likely and $\hat{q}_{D_{cal}}(\alpha)$ is the bias-adjusted $(1-\alpha)^{th}$ quantile of the APS conformity scores as defined for Eqn. 7.

Let $\hat{P}_{sum}(S) = \sum_{c_k \in S} \hat{p}_k$ denote the (fitted) probability mass within the prediction set S. Due to the unimodality of $\hat{p}$, it follows that one of the boundary labels c_u and c_{l+1} have the minimum probability among those included in the set $\hat{S}_{\mathcal{D},\alpha}(\mathbf{x})$. Without loss of generality, let us assume $\hat{p}_u$ is one of the minima (since the same argument can be applied for the case where $(l+1)$ is among the minima).

From the construction, we have, $\hat{P}_{sum}(\hat{S}_{\mathcal{D},\alpha}(\mathbf{x})) \geq \hat{q}_{D_{cal}}(\alpha)$ and $\hat{P}_{sum}(\hat{S}_{\mathcal{D},\alpha}(\mathbf{x}) \setminus \{c_u\}) < \hat{q}_{D_{cal}}(\alpha)$. Consider the sets $S_1 = \hat{S}_{\mathcal{D},\alpha}(\mathbf{x}) \bigcup \{c_{k^{skip}}\} \setminus \{c_u\}$ and $S_2 = S_1 \setminus \{c_{k^{min}}\}$ where k^{min} is the largest index satisfying $k^{min} = \underset{k|c_k \in S_1}{\arg\min}\, [\hat{p}_k]$. Since $\hat{p}_{k^{skip}} \geq \min(\hat{p}_{l+1}, \hat{p}_u)$, it follows that $\hat{P}_{sum}(S_1) \geq \hat{q}_{D_{cal}}(\alpha)$. Further, from the definition of j as the size of largest top k set with probability mass as defined in Eqn. 7, it follows that $\hat{P}_{sum}(S_2) < \hat{q}_{D_{cal}}(\alpha)$.

Therefore, the set S_1 is a valid APS prediction set as well with span $(k^{min} - l) < (u - l)$, which leads to a contradiction. Thus, the shortest span $\hat{S}_{D,\alpha}(\mathbf{x})$ has to be contiguous for this case as well. Thus, in both cases, there exists at least one solution, i.e., shortest span prediction set, which is contiguous for both the constructions.

For the case, where $\hat{p}$ is strictly unimodal, from the constructio Eqn.6 or 7, the prediction sets have to contain the top k most likely classes for some k which results in contiguity in case of strict unimodality. $\square$

Theorem 1. *For any $\mathbf{x} \in \mathcal{X}$, let $p_k(\mathbf{x}) = P_{Y|X}(Y = c_k|X = \mathbf{x})$ and $\hat{p}_k(\mathbf{x}) = \hat{P}_{Y|X}(Y = c_k|X = \mathbf{x})$ denote the true and fitted model class probabilities that are always unimodal. Let $\sigma_k(\mathbf{x}) = \sum_{k'=1}^{k} p_{k'}(\mathbf{x})$ and $\hat{\sigma}_k(\mathbf{x}) = \sum_{k'=1}^{k} \hat{p}_{k'}(\mathbf{x})$ denote the corresponding cumulative distribution functions. If $|\sigma_k(\mathbf{x}) - \hat{\sigma}_k(\mathbf{x})| \leq \delta,\ \ [k]_1^K$ for a constant δ, then for any $\alpha \in (0, 1]$, $\forall \mathbf{x} \in D_{test}$, the APS and oracle prediction sets from Eqn. 7 and Eqn. 4 satisfy $|\hat{S}_{D,\alpha}(\mathbf{x})| \leq |S^{oracle}_{\alpha-4\delta-\frac{1}{n+1}}(\mathbf{x})|$ where n is the size of the calibration set.*

Proof. To establish the result, we prove that the following two statements hold true under the assumption on the CDFs of $P_{Y|X}$ and $\hat{P}_{Y|X}$:

(a) $|\hat{S}_{D,\alpha}(\mathbf{x})| \leq |S^{Oracle}_{1-\hat{q}_{D_{cal}}(\alpha)-2\delta}(\mathbf{x})|$

(b) $|S^{oracle}_{1-\hat{q}_{D_{cal}}(\alpha)-2\delta}(\mathbf{x})| \leq |S^{oracle}_{\alpha-4\delta-\frac{1}{n+1}}|$

Part (a): From Eqn. 4 and Lemma 1, we observe that the unimodality of $\hat{p}(\mathbf{x})$ and $p(\mathbf{x})$ leads to the oracle prediction set being contiguous and also the existence of a contiguous APS prediction set. Since all the APS solution sets as per Eqn 7 have the same cardinality, we use $\hat{S}_{D,\alpha}(\mathbf{x})$ to denote the contiguous solution.

Let $\hat{S}_{D,\alpha}(\mathbf{x}) = \{c_{\hat{l}+1}, \cdots, c_{\hat{u}}\},\ \ 0 \leq \hat{l} < \hat{u} \leq K$ and $S^{oracle}_{1-\hat{q}_{D_{cal}}(\alpha)-2\delta}(\mathbf{x}) = \{c_{l^*+1}, \cdots, c_{u^*}\},\ \ 0 \leq l^* < u^* \leq K$. From the definition of the sets and the contiguity, we observe that the probability mass of $\hat{S}_{D,\alpha}(\mathbf{x})$ w.r.t. $\hat{p}$ equals $(\hat{\sigma}_{\hat{u}}(\mathbf{x}) - \hat{\sigma}_{\hat{l}}(\mathbf{x})) \geq \hat{q}_{D_{cal}}(\alpha)$ while that of $S^{oracle}_{1-\hat{q}_{D_{cal}}(\alpha)-2\delta}(\mathbf{x})$ w.r.t p equals $(\sigma_{u^*}(\mathbf{x}) - \sigma_{l^*}(\mathbf{x})) \geq 1 - (1 - \hat{q}_{D_{cal}}(\alpha) - 2\delta) = \hat{q}_{D_{cal}}(\alpha) + 2\delta$.

Using the divergence bound on the two CDFs, i.e., $|\sigma_k(\mathbf{x}) - \hat{\sigma}_k(\mathbf{x})| \leq \delta,\ \ [k]_1^K$, we have

$$\begin{aligned}
(\hat{\sigma}_{u^*}(\mathbf{x}) - \hat{\sigma}_{l^*}(\mathbf{x})) &\geq (\sigma_{u^*}(\mathbf{x}) - \delta) - ((\sigma_{l^*}(\mathbf{x}) + \delta) \\
&= \sigma_{u^*}(\mathbf{x}) - \sigma_{l^*}(\mathbf{x}) - 2\delta \\
&\geq \hat{q}_{D_{cal}}(\alpha) + 2\delta - 2\delta \\
&= \hat{q}_{D_{cal}}(\alpha).
\end{aligned}$$

Since $\hat{S}_{D,\alpha}(\mathbf{x})$ is the minimal contiguous set with probability mass greater than or or equal to $\hat{q}_{D_{cal}}(\alpha)$ as per $\hat{p}$ in Eqn 7, we have

$$|\hat{S}_{D,\alpha}(\mathbf{x})| = (\hat{u} - \hat{l}) \leq (u^* - l^*) = |S^{oracle}_{1-\hat{q}_{D_{cal}}(\alpha)-2\delta}(\mathbf{x})|.$$

Part (b): Denoting the minimal contiguous APS prediction set by $\hat{S}_{D,\alpha}(\mathbf{x})$ as before, we have $(\hat{\sigma}_{\hat{u}}(\mathbf{x}) - \hat{\sigma}_{\hat{l}}(\mathbf{x})) \geq \hat{q}_{D_{cal}}(\alpha)$. Considering the divergence bound on the two CDFs, i.e., $|\sigma_k(\mathbf{x}) - \hat{\sigma}_k(\mathbf{x})| \leq \delta,\ \ [k]_1^K$, we have $(\hat{\sigma}_{\hat{u}}(\mathbf{x}) - \hat{\sigma}_{\hat{l}}(\mathbf{x})) \leq (\sigma_{\hat{u}}(\mathbf{x}) + \delta) - ((\sigma_{\hat{l}}(\mathbf{x}) - \delta) = \sigma_{\hat{u}}(\mathbf{x}) - \sigma_{\hat{l}}(\mathbf{x}) + 2\delta$. Hence, for all $\mathbf{x}$, we have

$$\begin{aligned}
&(\hat{\sigma}_{\hat{u}}(\mathbf{x}) - \hat{\sigma}_{\hat{l}}(\mathbf{x})) \geq \hat{q}_{D_{cal}}(\alpha) \\
\Leftrightarrow\ \ &\sigma_{\hat{u}}(\mathbf{x}) - \sigma_{\hat{l}}(\mathbf{x}) + 2\delta \geq \hat{q}_{D_{cal}}(\alpha) \\
\Leftrightarrow\ \ &\sigma_{\hat{u}}(\mathbf{x}) - \sigma_{\hat{l}}(\mathbf{x}) \geq \hat{q}_{D_{cal}}(\alpha) - 2\delta
\end{aligned}$$

Since this holds for all $\mathbf{x}$, the marginal coverage $P[Y \in \hat{S}_{D,\alpha}(X)] \geq \hat{q}_{D_{cal}}(\alpha) - 2\delta$.

From Theorem 3, we also have an upper bound on the marginal coverage for test samples, i.e., $P[Y \in \hat{S}_{D,\alpha}(X)] \leq 1 - \alpha + \frac{1}{n+1}$ where n is the size of the calibration set.

Hence, we have

$$\begin{aligned}
& 1-\alpha+\frac{1}{n+1} \geq P[Y \in \hat{S}_{D,\alpha}(X)] \\
\Leftrightarrow \quad & 1-\alpha+\frac{1}{n+1} \geq \hat{q}_{D_{cal}}(\alpha)-2\delta \\
\Leftrightarrow \quad & 1-\hat{q}_{D_{cal}}(\alpha)-2\delta \geq \alpha-4\delta-\frac{1}{n+1}
\end{aligned}$$

From the above inequality and the definition of the oracle prediction set, we observe that

$$|S^{oracle}_{1-\hat{q}_{D_{cal}}(\alpha)-2\delta}(\mathbf{x})| \leq |S^{oracle}_{\alpha-4\delta-\frac{1}{n+1}}|.$$

Combining the results in part (a) and (b), we have

$$|\hat{S}_{D,\alpha}(\mathbf{x})| \leq |S^{oracle}_{\alpha-4\delta-\frac{1}{n+1}}|.$$

As the size of the calibration set increases, the term $\frac{1}{n+1}$ vanishes and as the divergence δ decreases, then the cardinality of the APS set converges to that of the oracle set.

□

Theorem 2. *Let $\eta : \mathcal{X} \rightarrow R^K$, $\phi : R \rightarrow R^+$ and $\psi^E : R \rightarrow R^-$ such that $\psi^E(r) = \psi^E(-r),\ \forall r \in R$ and its restriction to R^+ is a strictly monotonically decreasing bijective function. (a) Then, the model output constructed as per Eqn. 5 is always unimodal, i.e., $\hat{\mathrm{p}}(\mathbf{x}) \in \mathcal{U},\ \forall \mathbf{x} \in \mathcal{X}$. (b) Further, given any $\hat{\mathrm{p}}(\mathbf{x}) \in \mathcal{U}$ for $\mathbf{x} \in \mathcal{X}$, there exists a function $\eta(\mathbf{x}) \in R^K$ that satisfies Eqn. 5 if $\phi(\cdot)$ is surjective on R^+.*

Proof. We begin by restating the construction:

$$\begin{aligned}
&\eta(\mathbf{x}) = \mathbf{f}(\mathbf{x},\theta);\ \ v_1 = \eta_1(\mathbf{x});\ \ v_k = \phi(\eta_k(\mathbf{x})),\ \ [k]_2^K, \\
&r_1 = v_1;\ \ r_k = r_{k-1} + v_k,\ \ [k]_2^K;\ \ z_k = \psi^E(r_k);\ \ \hat{p}_k = \frac{\exp(z_k)}{\sum_{k=1}^K \exp(z_k)},\ \ [k]_1^K.
\end{aligned}$$

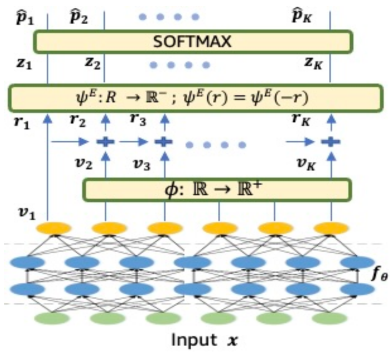

Figure 7: Construction of our DNN

Part a: Following the above construction, for any $\mathbf{x} \in \mathcal{X}$, since $\phi : R \rightarrow R^+$, the DNN output $v_k \geq 0,\ [k]_2^K$. Hence, the cumulative sum sequence $\mathbf{r}$ is non-decreasing, i.e., $r_1 \leq r_2 \leq \cdots \leq r_K$.

There can be 3 possible scenarios:

Scenario 1. $r_1 \leq r_2 \cdots \leq r_K \leq 0$: In this case, $[z_k = \psi^E(r_k)]_{k=1}^K$ is also a non-decreasing sequence and so is $[\hat{p}_k]_{k=1}^K$. Here $[\hat{p}_k]_{k=1}^K$ is unimodal sequence with mode at c_K.

Scenario 2. $0 \leq r_1 \leq r_2 \cdots <= r_K$: In this case, $[z_k = \psi^E(r_k)]_{k=1}^K$ is also a non-increasing sequence and so is $[\hat{p}_k]_{k=1}^K$. Here $[\hat{p}_k]_{k=1}^K$ is unimodal sequence with mode at c_1.

Scenario 3. $r_1 \leq r_2 \cdots \leq r_m \leq 0 \leq r_{m+1} \leq \cdots \leq r_K$ for some m. In this case, $[z_k = \psi^E(r_k)]_{k=1}^K$ is non-decreasing till m and non-increasing from $m+1$ onwards, which makes it unimodal. The mode is either m or $m+1$ or both depending on the magnitudes $|r_m|$ and $|r_{m+1}|$. The probability distribution $[\hat{p}_k]_{k=1}^K$ follows the same pattern and is unimodal as well.

Part b: Let us assume $\hat{p}(\mathbf{x}) \in \mathcal{U}$ is any arbitrary unimodal distribution conditioned on x with class probabilities $\hat{p}_k \leq \hat{p}_{k+1},\ [k]_1^{m-1}$ and $\hat{p}_k \geq \hat{p}_{k+1},\ [k]_m^K$, where m is the highest indexed (in case of multiple) mode of the unimodal distribution. We can then obtain $z_k = \log(\hat{p}_k),\ [k]_1^K$ and construct the sequence $r_k = (\psi^E)^{-1}(z_k)$ where $r_k \in R^-$ for $1 \leq k \leq (m-1)$ and $r_k \in R^+$ for $m \leq k \leq K$. Since $\psi : R^+ \to R^-$ is a strictly monotonically decreasing bijective function and ψ^E is it's even extension, the sequence $[r_k]_{k=1}^K$ is well-defined. Further, since $[z_k]_{k=1}^K$ is a unimodal sequence, $[r_k]_{k=1}^K$ is monotonically increasing with $r_{m-1} \leq 0 \leq r_m$. Then, we can obtain the vector $\mathbf{v}$ such that $v_k = r_k - r_{k-1} \geq 0,\ [k]_2^K$ and $v_1 = r_1$. When $\phi(\cdot)$ is a surjective function on R^+, we can define $\eta_k(\mathbf{x}) = (\phi)^{-1}(v_k),\ [k]_2^K$ and $\eta_1(\mathbf{x}) = v_1$. There will always be a valid $\eta(\mathbf{x})$, which ensures that construction can generate the original $\hat{p}(\mathbf{x})$. $\square$

B.1 APS Coverage guarantees

Theorem 3. *[APS [34]] If samples (x_i, y_i) $x_i \in \mathcal{X}, y_i \in \mathcal{Y}$ are exchangeable $\forall 1 \leq i \leq n$ and all samples from D_{train}, D_{cal} are invariant to permutations, and conformity scores are almost surely distinct, then APS algorithm gives tight marginal coverage given by:*

$$1 - \alpha \leq P[Y_{test} \in \hat{S}_{D,\alpha}(X_{test})] \leq 1 - \alpha + \frac{1}{|D_{cal}| + 1}$$

C Experiment Details

C.1 Benchmark Image Datasets and Implementation Details

We now provide a brief description of the four public datasets and the modeling details. For each of these datasets, we split the data into train, calibration, and test sets. We use calibration set to calibrate APS and report mean and standard deviation (std. error) on the test set across 5 independent splits. Note that for all experiments to avoid over-fitting, data augmentation, i.e., random horizontal flipping and random cropping for each training image, was applied in our experiments. The predictions was obtained with a central crop during testing. COPOC was implemented with $\phi = |x|$ and $\psi = -|x|$.

Age Estimation - Adience [13]: The task associated with this dataset is to predict the age for a given facial image. This dataset contains 26580 Flickr photos of 2284 subjects. The age is annotated with eight groups: $0-2, 4-6, 8-13, 15-20, 25-32, 38-43, 48-53$, and over 60 years old. From the nature of the class labels, it is evident that classes are not equally spaced categories. Hence, previous works which assumed it to be equi-spaced (SORD [12] for instance) are suboptimal. For feature extractor backbone, we use ImageNet pre-trained VGG-16 network since most competing methods [23, 12] used this model. For our usage we append single layer MLP with last layer configured to output unimodal distribution as described in sec. 4.2. We trained models for 50 epochs with a batch size of 64. For optimization, Adam optimizer was utilized with a learning rate of 0.0001, with decay rate of 0.2.
Historical Colour Image Dating - HCI [29]: The historical color image dataset is collected for the task of estimating the age of historical color photos. Each image is annotated with its associated decade, where five decades from the 1930s to 1970s are considered. There are 265 images for each category. Following [23] we utilized VGG-16 as the backbone, which was initialized with the ImageNet pre-trained weights for a fair comparison. We trained models for 50 epochs with Adam optimizer with a learning rate of 0.0001, with decay rate of 0.2. For COPOC, we append single layer MLP with last layer configured to output unimodal distribution as described in sec. 4.2.

Retina-MNIST [42]: RetinaMNIST is based on the DeepDRiD24 challenge, which provides a dataset of 1600 retina fundus images. The task is ordinal classification for 5-level grading of diabetic retinopathy severity. We use a similar feature extractor network as used in [42] along with a final unimodality constrained layer at end. The network was trained with same settings as [42].

Image Aesthetic Estimation [36]: The Aesthetics dataset consists of 15687 Flickr image belonging to four different nominal categories: animals, urban, people, and nature. All The pictures are annotated by 5 different graders in 5 aesthetic categories in an orderly manner: 1) "unacceptable" pictures with extremely low quality, 2) "flawed" low quality images (slightly blurred,over/underexposed), and with no artistic value; 3) "ordinary" images without technical flaws (well framed, in focus), but no artistic value; 4) "professional" images (flawless framing,lightning),and 5) "exceptional", very appealing images, showing outstanding quality. The ground truth label for each image is set to be the median among all of its gradings. Following [23, 12] we use ImageNet pre-trained VGG-16 as the backbone for feature extraction. For our usage, we append single layer MLP with last layer configured to output unimodal distribution as described in sec. 4.2. We only report aggregate metric across all the categories for this data.

Table 4: Results on Image Benchmark Datasets: Mean and std. error is reported for 5 trials. Best mean results bolded.

		MAE	Acc@1	Acc@2	Acc@3	|PS|	CV%
HCI	V-CE	0.68 ± 0.03	54.3 ± 2.6	75.3 ± 3.1	88.9 ± 1.6	3.28 ± 0.14	24.4 ± 1.2
	POE	0.66 ± 0.05	56.5 ± 1.8	76.5 ± 2.5	89.0 ± 2.1	3.1 ± 0.18	9.8 ± 1.2
	SORD	0.65 ± 0.06	56.2 ± 2.8	77.1 ± 2.9	89.8 ± 2.6	2.96 ± 0.19	2.7 ± 1.1
	AVDL	$\mathbf{0.64 \pm 0.08}$	$\mathbf{56.8 \pm 1.5}$	77.9 ± 2.4	89.8 ± 1.05	2.98 ± 0.11	2.1 ± 1.4
	Binomial	0.68 ± 0.05	54.5 ± 1.2	75.8 ± 2.6	88.8 ± 1.8	3.01 ± 0.16	0
	Binomial-temp	0.66 ± 0.04	55.5 ± 1.8	78 ± 2.2	90.1 ± 2.1	2.90 ± 0.11	0
	Uni-loss	0.67 ± 0.09	54.5 ± 3.1	74.8 ± 2.5	88.1 ± 2.5	3.05 ± 0.38	5.1 ± 1.9
	COPOC	0.65 ± 0.04	56.1 ± 2.0	$\mathbf{79.8 \pm 1.6}$	$\mathbf{91.7 \pm 2.8}$	$\mathbf{2.66 \pm 0.13}$	0
Adience	V-CE	0.57 ± 0.07	58.1 ± 1.6	80.8 ± 1.6	91.4 ± 2.3	4.82 ± 0.24	21.4 ± 2.2
	POE	$\mathbf{0.48 \pm 0.05}$	60.5 ± 1.5	84.1 ± 2.0	93.9 ± 2.3	4.16 ± 0.18	12.8 ± 1.2
	SORD	$\mathbf{0.48 \pm 0.06}$	59.9 ± 3.8	85.2 ± 2.9	94.3 ± 1.6	2.86 ± 0.09	3.7 ± 1.1
	AVDL	0.49 ± 0.03	60.1 ± 2.5	85.3 ± 3.1	94.0 ± 1.1	2.95 ± 0.15	4.1 ± 0.9
	Binomial	0.5 ± 0.04	60.0 ± 1.2	86 ± 1.8	95.4 ± 1.9	2.5 ± 0.06	0
	Binomial-temp	$\mathbf{0.48 \pm 0.04}$	60.5 ± 2.1	$\mathbf{86.4 \pm 1.2}$	95.6 ± 1.3	2.45 ± 0.05	0
	Uni-loss	0.64 ± 0.14	51.5 ± 7.9	80.8 ± 5.8	89.4 ± 3.5	3.14 ± 0.26	8.3 ± 2.3
	COPOC	0.49 ± 0.04	$\mathbf{61.0 \pm 1.9}$	86 ± 1.5	$\mathbf{96.1 \pm 2.2}$	$\mathbf{2.26 \pm .06}$	0
Aesthetic	V-CE	0.29 ± 0.01	71.4 ± 1.6	94.6 ± 2.0	97.8 ± 0.8	1.96 ± 0.2	7.9 ± 0.2
	POE	$\mathbf{0.28 \pm 0.05}$	72.1 ± 1.5	94.1 ± 1.1	98.0 ± 0.1	1.85 ± 0.11	7.85 ± 0.9
	SORD	0.29 ± 0.02	72.0 ± 1.7	95.2 ± 1.9	98.3 ± 0.2	1.78 ± 0.09	0
	AVDL	$\mathbf{0.28 \pm 0.03}$	$\mathbf{72.2 \pm 1.5}$	95.2 ± 1.8	98.5 ± 0.1	1.75 ± 0.05	0.3 ± 0.1
	Binomial	0.31 ± 0.01	69.5 ± 0.7	93.1 ± 2.8	96.0 ± 0.9	1.83 ± 0.06	0
	Binomial-temp	0.32 ± 0.04	69 ± 1.7	93.0 ± 1.6	96.2 ± 0.1	1.89 ± 0.09	0
	Uni-loss	0.37 ± 0.14	66.8 ± 5.0	92.0 ± 3.8	97.4 ± 1.5	1.94 ± 0.24	2.1 ± 0.8
	COPOC	$\mathbf{0.28 \pm 0.04}$	72.0 ± 1.3	$\mathbf{95.9 \pm 1.0}$	$\mathbf{99.0 \pm 0.2}$	$\mathbf{1.70 \pm .06}$	0
Retina-MNIST	V-CE	0.73 ± 0.02	52.2 ± 0.6	72.2 ± 0.1	86.0 ± 0.5	3.6 ± 0.08	9.8 ± 2.4
	POE	0.73 ± 0.02	52.4 ± 0.4	$\mathbf{72.5 \pm 0.6}$	86.4 ± 0.8	3.4 ± 0.05	6.4 ± 2.8
	SORD	0.71 ± 0.01	$\mathbf{53.5 \pm 0.3}$	70.5 ± 0.6	84.5 ± 0.9	3.2 ± 0.03	3.9 ± 1.1
	AVDL	0.72 ± 0.02	53.0 ± 0.2	71.0 ± 0.4	84.6 ± 0.9	3.24 ± 0.04	3.8 ± 1.2
	Binomial	0.71 ± 0.01	52.7 ± 0.2	69.7 ± 0.6	83.7 ± 0.8	3.33 ± 0.02	0
	Binomial-temp	$\mathbf{0.70 \pm 0.02}$	53.0 ± 0.2	70.5 ± 0.5	84.0 ± 0.4	3.3 ± 0.02	0
	Uni-loss	0.74 ± 0.05	52.0 ± 1.1	$\mathbf{72.5 \pm 0.6}$	84.5 ± 1.5	3.25 ± 0.1	4.2 ± 1.1
	COPOC	0.71 ± 0.01	$\mathbf{53.5 \pm 0.2}$	$\mathbf{72.5 \pm 0.6}$	$\mathbf{87.0 \pm 0.3}$	$\mathbf{3.03 \pm 0.01}$	0

Result Discussion : We highlight the key takeaways from Table 4

1. COPOC performs at par with SOTA baselines in terms of MAE and $Acc@1$.

2. Benefit of COPOC comes with improved gains in $Acc@2$ and $Acc@3$. Apart from COPOC, there is no single method that performs consistently across the 4 datasets in terms of these metrics. For instance in *HCI* and *Adience*, *Binomial-temp* comes closest to *COPOC*, but on *Aesthetic*, both

variants of Binomial severely under-perform whereas *AVDL* and *SORD* perform quite well and comes the closest to *COPOC*. In contrast, on *Retina-MNIST*, non-parametric models such as *V-CE*,*POE*, *Uni-loss* have $Acc@k$ close to *COPOC* and beat other parametric models significantly. This shows that parametric distribution assumption in any underlying model fits the data well when the data is actually drawn from a similar distribution. Since most methods depend largely on the validity of the assumptions, the relatively unconstrained parameter-free nature of COPOC is more robust and allows it to consistently outperform across datasets.

3. There is a strong correlation between $CV\%$ and PS size. This is expected because higher $CV\%$ indicates more number of cases for which we had to predict a minimal contiguous super set, thus inflating the size of PS. Better unimodal fit by underlying model is bound to have lesser $CV\%$ and and thus, shorter sets. Hence, COPOC again outperforms all other baselines across datasets in term of |PS| consistently. Although *Binomial* model variants has 0 $CV\%$ due to it's construction, it still produces larger sets than *COPOC* as seen in *HCI* and *Adience*. This can be because COPOC results in better unimodal fit which is also idicated by higher $Acc@K$.

4. Enforcing unimodality in training scheme in terms of soft-labels ($SORD$,$AVDL$) or in loss function (*Uni-loss*) or in embedding space (POE) does not necessarily translate to a unimodal distribution in test samples which is indicated by high $CV\%$.

5. Although $V - CE$ in principle should have been able to model any underlying distribution, on high dimensional real-world datasets it fails miserably. This shows the need for injecting prior "bias" into training network like COPOC which aids the model in reaching the optima.

6. *Uni-loss* has issues with model convergence as it shows high variance across metrics for all datasets. This could be because its sensitive to λ hyperparamter that control the weightage between unimodality and mean-variance component of its loss function which is difficult to tune.

7. Datasets with higher accuracy results in shorter PS size in general, which is expected. For instance *Aesthetic* has lower PS size across methods compared to *HCI* or *Retina-MNIST* both having same number of class labels.

C.2 Implementation details of experiments on synthetic Data

For all the results on synthetic datasets presented in Sec.5.3, we employ same DNN network across all the methods for fair comparison. To be precise, we use 6 layer DNN architecture having 128 hidden dimensions with a dropout of 0.2. We use the same training paradigm as before – Adam optimizer with learning rate of 0.001 and batch size of 512 trained for 500 epochs ensuring convergence. We divide the data into 70% train and 30% test splits. We train our model 10 times for each independent split of the data. For each test set, we again randomly split into calibration for APS and evaluate |PS| on final-test and repeat this 100 times to ensure convergence of PS. We use $\phi = |x|$ and $\psi = -|x|$ for COPOC. We report mean and standard error in Table 2.

C.3 Ablation study on the choice of $\phi(\cdot)$ and $\psi^E(\cdot)$ for COPOC

Although there can be many possible choices for $\phi(\cdot)$ and $\psi^E(\cdot)$ in the COPOC construction Eqn. 5, in practice not all choices leads to good model convergence. In this section, we perform a comparison of few common choices and present results in Table 5. We train the model for synthetic data D4 as described in Sec. 5.3. We train the model using CE loss with same model capacity and training paradigm as described in Appendix C.2. We report train CE loss and since we have access to true underlying distribution for D4 we report KL. Div. to measure goodness of model fit. Below we summarise few observations:

1. $\phi = ReLU$ maps most of $[v_k]_2^K$ to zeroes which results in flat probability distribution for most of the data points while $\phi = Softplus$ instead maps most $[v_k]_2^K$ to very small values which again results in almost flat distribution for most points. With $\phi = x^2$ we observed unusually large values for $[v_k]_2^K$ resulting in unstable training. $\phi = |x|$ gives a good balance as each $[\eta_k]_2^K$ gets linearly mapped to $[v_k]_2^K$.
2. $\psi = -|x|^2$ tends to over-emphasize higher probability classes in the model fitting while $\psi = -|x|^{0.5}$ under-emphasizes them. Again since $\psi = -|x|$ does a linear transformation of r_k on either side of the origin it gives a good balanced estimate of z_k.

Table 5: Ablation study on implementation choice of $\phi(\cdot)$ and $\psi(\cdot)$ for COPOC. We report mean results across 10 trials.

	Train loss	KL Div.
$\phi = ReLU, \psi = -\|x\|$	3.89	0.24
$\phi = Softplus, \psi = -\|x\|$	3.11	0.19
$\phi = x^2, \psi = -\|x\|$	3.48	0.2
$\phi = \|x\|, \psi = -\|x\|$	$\mathbf{1.64}$	$\mathbf{0.04}$
$\phi = \|x\|, \psi = -x^2$	2.20	0.13
$\phi = \|x\|, \psi = -\|x\|^{0.5}$	1.89	0.1

References

[1] A. Angelopoulos, S. Bates, J. Malik, and M. I. Jordan. Uncertainty sets for image classifiers using conformal prediction. *arXiv preprint arXiv:2009.14193*, 2020.

[2] A. N. Angelopoulos and S. Bates. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv preprint arXiv:2107.07511*, 2021.

[3] A. N. Angelopoulos, A. P. Kohli, S. Bates, M. Jordan, J. Malik, T. Alshaabi, S. Upadhyayula, and Y. Romano. Image-to-image regression with distribution-free uncertainty quantification and applications in imaging. In *International Conference on Machine Learning*, pages 717–730. PMLR, 2022.

[4] C. Beckham and C. Pal. Unimodal probability distributions for deep ordinal classification. In *International Conference on Machine Learning*, pages 411–419. PMLR, 2017.

[5] S. Belharbi, I. B. Ayed, L. McCaffrey, and E. Granger. Non-parametric uni-modality constraints for deep ordinal classification. *arXiv preprint arXiv:1911.10720*, 2019.

[6] A. Berg, M. Oskarsson, and M. O'Connor. Deep ordinal regression with label diversity. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 2740–2747. IEEE, 2021.

[7] R. Can Malli, M. Aygun, and H. Kemal Ekenel. Apparent age estimation using ensemble of deep learning models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 9–16, 2016.

[8] W. Cao, V. Mirjalili, and S. Raschka. Rank consistent ordinal regression for neural networks with application to age estimation. *Pattern Recognition Letters*, 140:325–331, 2020.

[9] M. Cauchois, S. Gupta, and J. C. Duchi. Knowing what you know: valid and validated confidence sets in multiclass and multilabel prediction. *The Journal of Machine Learning Research*, 22(1):3681–3722, 2021.

[10] J. F. P. da Costa, H. Alonso, and J. S. Cardoso. The unimodal model for the classification of ordinal data. *Neural Networks*, 21(1):78–91, 2008.

[11] A. Daniely, N. Linial, and S. Shalev-Shwartz. From average case complexity to improper learning complexity. In *Proceedings of the forty-sixth annual ACM symposium on Theory of computing*, pages 441–448, 2014.

[12] R. Diaz and A. Marathe. Soft labels for ordinal regression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4738–4747, 2019.

[13] E. Eidinger, R. Enbar, and T. Hassner. Age and gender estimation of unfiltered faces. *IEEE Transactions on information forensics and security*, 9(12):2170–2179, 2014.

[14] A. Fisch, T. Schuster, T. Jaakkola, and R. Barzilay. Few-shot conformal prediction with auxiliary tasks. In *International Conference on Machine Learning*, pages 3329–3339. PMLR, 2021.

[15] B.-B. Gao, C. Xing, C.-W. Xie, J. Wu, and X. Geng. Deep label distribution learning with label ambiguity. *IEEE Transactions on Image Processing*, 26(6):2825–2838, 2017.

[16] X. Geng. Label distribution learning. *IEEE Transactions on Knowledge and Data Engineering*, 28(7):1734–1748, 2016.

[17] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR, 2017.

[18] G. Guo, Y. Fu, C. R. Dyer, and T. S. Huang. Image-based human age estimation by manifold learning and locally adjusted robust regression. *IEEE Transactions on Image Processing*, 17(7):1178–1188, 2008.

[19] Y. Hechtlinger, B. Póczos, and L. Wasserman. Cautious deep learning. *arXiv preprint arXiv:1805.09460*, 2018.

[20] Z. Huo, X. Yang, C. Xing, Y. Zhou, P. Hou, J. Lv, and X. Geng. Deep age distribution learning for apparent age estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 17–24, 2016.

[21] J. Lei, M. G'Sell, A. Rinaldo, R. J. Tibshirani, and L. Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523):1094–1111, 2018.

[22] Q. Li, J. Wang, Z. Yao, Y. Li, P. Yang, J. Yan, C. Wang, and S. Pu. Unimodal-concentrated loss: Fully adaptive label distribution learning for ordinal regression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20513–20522, 2022.

[23] W. Li, X. Huang, J. Lu, J. Feng, and J. Zhou. Learning probabilistic ordinal embeddings for uncertainty-aware regression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13896–13905, 2021.

[24] X. Liu, X. Han, Y. Qiao, Y. Ge, S. Li, and J. Lu. Unimodal-uniform constrained wasserstein training for medical diagnosis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019.

[25] Y. Lu and J. Lu. A universal approximation theorem of deep neural networks for expressing probability distributions. *Advances in neural information processing systems*, 33:3094–3105, 2020.

[26] B. Neyshabur, R. Tomioka, and N. Srebro. In search of the real inductive bias: On the role of implicit regularization in deep learning. *arXiv preprint arXiv:1412.6614*, 2014.

[27] Z. Niu, M. Zhou, L. Wang, X. Gao, and G. Hua. Ordinal regression with multiple output cnn for age estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4920–4928, 2016.

[28] J. Nixon, M. W. Dusenberry, L. Zhang, G. Jerfel, and D. Tran. Measuring calibration in deep learning. In *CVPR workshops*, volume 2, 2019.

[29] F. Palermo, J. Hays, and A. A. Efros. Dating historical color images. In *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part VI 12*, pages 499–512. Springer, 2012.

[30] H. Papadopoulos, K. Proedrou, V. Vovk, and A. Gammerman. Inductive confidence machines for regression. In *Machine Learning: ECML 2002: 13th European Conference on Machine Learning Helsinki, Finland, August 19–23, 2002 Proceedings 13*, pages 345–356. Springer, 2002.

[31] J. Platt et al. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10(3):61–74, 1999.

[32] J. D. Rennie and N. Srebro. Loss functions for preference levels: Regression with discrete ordered labels. In *Proceedings of the IJCAI multidisciplinary workshop on advances in preference handling*, volume 1. AAAI Press, Menlo Park, CA, 2005.

[33] Y. Romano, E. Patterson, and E. Candes. Conformalized quantile regression. *Advances in neural information processing systems*, 32, 2019.

[34] Y. Romano, M. Sesia, and E. Candes. Classification with valid and adaptive coverage. *Advances in Neural Information Processing Systems*, 33:3581–3591, 2020.

[35] M. Sadinle, J. Lei, and L. Wasserman. Least ambiguous set-valued classifiers with bounded error levels. *Journal of the American Statistical Association*, 114(525):223–234, 2019.

[36] R. Schifanella, M. Redi, and L. M. Aiello. An image is worth more than a thousand favorites: Surfacing the hidden beauty of flickr pictures. In *Proceedings of the international AAAI conference on web and social media*, volume 9, pages 397–406, 2015.

[37] K. Stankeviciute, A. M Alaa, and M. van der Schaar. Conformal time-series forecasting. *Advances in Neural Information Processing Systems*, 34:6216–6228, 2021.

[38] T. Vishnu, P. Malhotra, L. Vig, G. Shroff, et al. Data-driven prognostics with predictive uncertainty estimation using ensemble of deep ordinal regression models. *International Journal of Prognostics and Health Management*, 10(4), 2019.

[39] V. Vovk, A. Gammerman, and C. Saunders. Machine-learning applications of algorithmic randomness. 1999.

[40] V. Vovk, A. Gammerman, and G. Shafer. *Algorithmic learning in a random world*, volume 29. Springer, 2005.

[41] X. Wen, B. Li, H. Guo, Z. Liu, G. Hu, M. Tang, and J. Wang. Adaptive variance based label distribution learning for facial age estimation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIII 16*, pages 379–395. Springer, 2020.

[42] J. Yang, R. Shi, D. Wei, Z. Liu, L. Zhao, B. Ke, H. Pfister, and B. Ni. Medmnist v2-a large-scale lightweight benchmark for 2d and 3d biomedical image classification. *Scientific Data*, 10(1):41, 2023.

[43] D. Yi, Z. Lei, and S. Z. Li. Age estimation by multi-scale convolutional network. In *Computer Vision–ACCV 2014: 12th Asian Conference on Computer Vision, Singapore, Singapore, November 1-5, 2014, Revised Selected Papers, Part III 12*, pages 144–158. Springer, 2015.