An Information Theoretic Perspective on Conformal Prediction

Kim Choeun

February 25, 2025

Seoul National University

Outline

1. Introduction
2. Background
3. Information Theory Applied to Conformal Prediction
4. Conformal Training
5. Experiments

Introduction

Introduction

- in this work, we take a closer look at conformal prediction through the lens of information theory
- proved conformal prediction can be used to bound $H(Y|X)$ in three different ways : DPI bound, Fano bound, model-based Fano bound
- showed the upper bounds serve as principled training objectives to learn classifiers that are more amenable to SCP
- validate that both these applications of our theoretical results lead to better predictive efficiency, i.e., narrower and, consequently, more informative prediction sets

Background

Conformal Prediction

- Conformal Prediction (CP) is a framework that provides prediction sets with finite-sample guarantees under minimal distribution-free assumptions
- Given a set of n data points $(X_i, Y_i) \in \mathcal{X} \times \mathcal{Y},\ i = 1, \ldots, n$ drawn from some joint distribution P_{XY}, CP allows us to construct sets $\mathcal{C}(X) \in \mathcal{Y}$ such that

$$\mathbb{P}\left(Y_{test} \in \mathcal{C}\left(X_{test}\right)\right) \geq 1 - \alpha \text{ where } \left(X_{test}, Y_{test}\right) \sim P_{XY}$$

Conformal Prediction

Split Conformal Predcition (SCP)

- can leverage any pre-trained model $f : \mathcal{X} \to \mathcal{Y}$ in the construction of prediction sets
- aforesaid n data points constitute a calibration data set $\mathcal{D}_{cal}$, which must be disjoint from the training data set used to fit the predictive model f
- Procedure
  1. define a nonconformity score function $s_f : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ which captures the magnitude of the prediction error at a given data point (the higher the score, the higher the disagreement between input x and prediction y)
  2. evaluate the score function at every $(X_i, Y_i) \in \mathcal{D}_{cal}$ to get a collection of scores $\{S_i = s_f(X_i, Y_i)\}_{i=1}^n$
  3. construct prediction set $\mathcal{C}(X_{test})$ as $\mathcal{C}(X_{test}) = \{y \in Y : s_f(X_{test}, y) \leq \mathit{Quantile}(1-\alpha; \{S_i\}_{i=1}^n \cup \{\infty\})\}$ where $\mathit{Quantile}(1-\alpha; \{S_i\}_{i=1}^n)$ is the level $1-\alpha$ quantile of the empirical distribution defined by $\{S_i\}_{i=1}^n$

Conformal Prediction

Theorem1

If $\{X_i, Y_i\}_{i=1}^n$ are i.i.d. (or only exchangeable), then for a new i.i.d. draw (X_{test}, Y_{test}), and for any $\alpha \in (0, 1)$ and for any score function such that $\{S_i\}_{i=1}^n$ are almost surely distinct, then $\mathcal{C}\left(X_{test}\right)$ as defined above satisfies

$$1 - \alpha \leq \mathbb{P}\left(Y_{test} \in \mathcal{C}\left(X_{test}\right)\right) \leq 1 - \alpha_n, \ \textit{where } \alpha_n = \alpha - \frac{1}{n+1}$$

We would also like our prediction sets to be as narrow as possible, and that is why CP methods are often compared in terms of their (empirical) inefficiency, i.e., the average prediction set size $\frac{1}{|\mathcal{D}_{test}|}\sum_{x \in \mathcal{D}_{test}} |\mathcal{C}\left(x\right)|$ for some test data set $\mathcal{D}_{test}$.

Information Theory Applied to Conformal Prediction

DPI

DPI for f-divergence

For any two probability measures P_X and Q_X defined on a space $\mathcal{X}$, and any map $W_{Y|X}$, which maps (P_X, Q_X) to (P_Y, Q_Y), we have

$$D_f\left(P_X||Q_X\right) \geq D_f\left(P_Y||Q_Y\right)$$

where f is a convex function with $f\left(1\right) = 0$, and the f-divergence between two probability measures P and Q is defined as

$$D_f\left(P||Q\right) := \mathbb{E}_Q\left[f\left(\frac{dP}{dQ}\right)\right]$$

DPI bound

DPI for CP

Consider any conformal prediction method with the prediction set $\mathcal{C}(x)$ with the following finite sample guarantee:

$$1-\alpha \leq \mathbb{P}(Y \in \mathcal{C}(x)) \leq 1-\alpha+\frac{1}{n+1}$$

for any $\alpha \in (0, 0.5)$.

For any arbitrary conditional distribution $Q_{Y|X}$, the true conditional distribution $P_{Y|X}$ and the input measure P_X, define the following two measures $Q := P_X Q_{Y|X}$ and $P := P_X P_{Y|X}$. We have for any $\alpha \in (0, 0.5)$,

$$\begin{aligned} H(Y|X) \leq & h_b(\alpha) + \left(1-\alpha+\frac{1}{n+1}\right) logQ(Y \in \mathcal{C}(x)) \\ & + \alpha logQ(Y \notin \mathcal{C}(x)) \mathbb{E}_{P_{XY}}\left[logQ_{Y|X}\right] \end{aligned}$$

with $h_b(\alpha) = -\alpha log(\alpha) - (1-\alpha) log(1-\alpha)$.

DPI bound

Since the term $Q\left(Y \in \mathcal{C}\left(X\right)\right)$ appears inside a log, an empirical estimate $\hat{Q}\left(Y \in \mathcal{C}\left(X\right)\right)$ would result in a lower bound and would be biased.

DPI for CP

Based on the empirical Bernstein inequality, with probability $1-\delta$, we have

$$\Delta_{\delta}\left(\mathbf{Z}, n\right) := \sqrt{\frac{2V_n\left(\mathbf{Z}\right) log\left(2/\delta\right)}{n}} + \frac{7 log\left(2/\delta\right)}{3\left(n-1\right)}$$

$$Q\left(Y \in \mathcal{C}\left(X\right)\right) \leq \hat{Q}\left(Y \in \mathcal{C}\left(X\right)\right) + \Delta_{\delta}\left(\mathbf{Z}, n\right) := \tilde{Q}\left(Y \in \mathcal{C}\left(X\right)\right),$$

$$Q\left(Y \notin \mathcal{C}\left(X\right)\right) \leq \hat{Q}\left(Y \notin \mathcal{C}\left(X\right)\right) + \Delta_{\delta}\left(\mathbf{Z}, n\right) := \tilde{Q}\left(Y \notin \mathcal{C}\left(X\right)\right),$$

with $V_n\left(\mathbf{Z}\right)$ the empirical variance of $\mathbf{Z} = \left(Z_1, \ldots, Z_n\right), Z_i = Q\left(y_i \in \mathcal{C}\left(x_i\right)\right)$. Using these bounds, we get the following inequality with probability $1-\delta$:

$$\begin{aligned} H\left(Y|X\right) \leq & h_b\left(\alpha\right) + \left(1-\alpha\right) log \tilde{Q}\left(Y \in \mathcal{C}\left(X\right)\right) + \\ & \alpha_n log \tilde{Q}\left(Y \notin \mathcal{C}\left(X\right)\right) - \mathbb{E}_P\left[log Q_{Y|X}\right] \end{aligned}$$

MB Fano bound

Model-Based Fano Bound

Consider any conformal prediction method satisfying the upper and lower bounds of Theorem1 for $\alpha \in (0, 0.5)$. Then, for the true distribution P, and for any probability distribution Q, we have

$$\begin{aligned} H\left(Y|X\right) \leq & h_b\left(\alpha\right) + \alpha \mathbb{E}_{P_{Y,X,\mathcal{D}_{cal}|Y\notin\mathcal{C}(X)}}\left[-logQ_{Y|X,\mathcal{C}(X),Y\notin\mathcal{C}(X)}\right] \\ & + \left(1-\alpha_n\right)\mathbb{E}_{P_{Y,X,\mathcal{D}_{cal}|Y\in\mathcal{C}(X)}}\left[-logQ_{Y|X,\mathcal{C}(X),Y\in\mathcal{C}(X)}\right] \end{aligned}$$

A good choice for Q is the predictive model itself, and that is why we refer to the bound above as Model-Based Fano bound.

Simple Fano bound

Simple Fano Bound

Consider any conformal prediction method satisfying the upper and lower bounds of Theorem1 for $\alpha \in (0, 0.5)$. Then, for the true distribution P we have

$$\begin{aligned} H(Y|X) \leq & h_b(\alpha) + \alpha \mathbb{E}_{P_{Y,X,\mathcal{D}_{cal}|Y \notin \mathcal{C}(X)}} \left[\, log \left(|\mathcal{Y}| - |\mathcal{C}(X)| \right) \, \right] \\ & + (1 - \alpha_n) \mathbb{E}_{P_{Y,X,\mathcal{D}_{cal}|Y \in \mathcal{C}(X)}} \left[\, log |\mathcal{C}(X)| \, \right] \end{aligned}$$

The proof follows directly from MB Fano bound by replacing Q with the uniform distribution.

Conformal Training

Conformal Training

- Although SCP is applicable to any pretrained ML model as a post-processing step, the overall performance of any CP method (commonly its inefficiency) is highly dependent on the underlying model itself.
- Therefore, previous works have proposed to take CP into account already during model training and directly optimize for low predictive inefficiency.
- In particular, ConfTr splits each training batch $\mathcal{B}$ into calibration $\mathcal{B}_{cal}$ and test $\mathcal{B}_{test}$ to simulate the SCP process for each gradient update of model f and minimize the following size loss

$$log \mathbb{E}\left[\left|\mathcal{C}_f\left(X\right)\right|\right] \approx log\left(1/|\mathcal{B}_{test}| \sum_{x \in \mathcal{B}_{test}} \left|\mathcal{C}_f\left(x\right)\right|\right)$$

Conformal Training

- Since SCP involves step functions, ConfTr introduced a couple of relaxations to recover a differentiable objective
    - the computation of quantiles is relaxed via differentiable sorting operators
    - the thresholding operation is replaced by smooth assignments of labels to prediction sets via the logistic sigmoid
- DPI, MB Fano and simple Fano bounds can be made differentiable in the same way and thus can serve as proper loss functions for conformal training.

Minimizing the UBs

- $H(Y|X)$ captures the underlying uncertainty under the true labelling distribution $P_{Y|X}$.
  - Minimizing the bounds, we can hope to push the model f closer to the true distribution, which is known to achieve minimal inefficiency.
  - Interestingly, the cross-entropy loss also bounds $H(Y|X)$ and thus can be motivated as a conformal training objective from the same angle.
  - In that regard, the DPI bound is particularly advantageous as it is provably tighter than the cross-entropy.

Minimizing the UBs

- We can connect the simple Fano bound to the size loss $\mathbb{E}\left[|\mathcal{C}\left(X\right)|\right]$.
- Applying Jensen's inequality and the fact that $log\left(|\mathcal{Y}| - |\mathcal{C}\left(X\right)|\right) \leq log|\mathcal{Y}|$ on the simple Fano bound, we obtain further UB.

$$H\left(Y|X\right) \leq \lambda_{\alpha} + \left(1 - \alpha_{n}\right) log\mathbb{E}\left[|\mathcal{C}\left(X\right)|\right]$$

- Therefore, we ground ConfTr as minimizing an UB to the true conditional entropy that is looser than our UBs for an appropriate choice for Q.

Experiments

Setup

- We test effectiveness of UBs as objectives for conformal training in five data sets : MNIST, Fashion-MNIST, EMNIST, CIFAR10 and CIFAR100.
- We follow a similar optimization procedure and experimental setup to that of ConfTr, but with the key differences that we learn the classifiers from scratch in all cases (without the need of pretrained CIFAR models).
- For each dataset, we use the default train and test splits but transfer 10% of the training data to the test dataset. We train the classifiers only on the remaining 90% of the training data and, at test time, run SCP with 10 different calibration/test splits by randomly splitting the enlarged test dataset.

Result

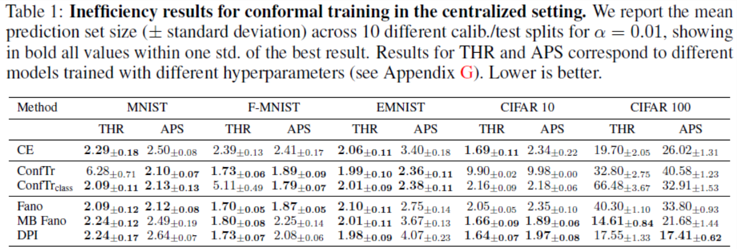


Table 1: **Inefficiency results for conformal training in the centralized setting.** We report the mean prediction set size (± standard deviation) across 10 different calib./test splits for $\alpha = 0.01$, showing in bold all values within one std. of the best result. Results for THR and APS correspond to different models trained with different hyperparameters (see Appendix G). Lower is better.

Method	MNIST		F-MNIST		EMNIST		CIFAR 10		CIFAR 100	
	THR	APS	THR	APS	THR	APS	THR	APS	THR	APS
CE	$\mathbf{2.29_{\pm 0.18}}$	$2.50_{\pm 0.08}$	$2.39_{\pm 0.13}$	$2.41_{\pm 0.17}$	$\mathbf{2.06_{\pm 0.11}}$	$3.40_{\pm 0.18}$	$\mathbf{1.69_{\pm 0.11}}$	$2.34_{\pm 0.22}$	$19.70_{\pm 2.05}$	$26.02_{\pm 1.31}$
ConfTr	$6.28_{\pm 0.71}$	$\mathbf{2.10_{\pm 0.07}}$	$\mathbf{1.73_{\pm 0.06}}$	$\mathbf{1.89_{\pm 0.09}}$	$\mathbf{1.99_{\pm 0.10}}$	$\mathbf{2.36_{\pm 0.11}}$	$9.90_{\pm 0.02}$	$9.98_{\pm 0.00}$	$32.80_{\pm 2.75}$	$40.58_{\pm 1.23}$
$\text{ConfTr}_{\text{class}}$	$\mathbf{2.09_{\pm 0.11}}$	$\mathbf{2.13_{\pm 0.13}}$	$5.11_{\pm 0.49}$	$\mathbf{1.79_{\pm 0.07}}$	$\mathbf{2.01_{\pm 0.09}}$	$\mathbf{2.38_{\pm 0.11}}$	$2.16_{\pm 0.09}$	$2.18_{\pm 0.06}$	$66.48_{\pm 3.67}$	$32.91_{\pm 1.53}$
Fano	$\mathbf{2.09_{\pm 0.12}}$	$\mathbf{2.12_{\pm 0.08}}$	$\mathbf{1.70_{\pm 0.05}}$	$\mathbf{1.87_{\pm 0.05}}$	$\mathbf{2.10_{\pm 0.11}}$	$2.75_{\pm 0.14}$	$2.05_{\pm 0.05}$	$2.35_{\pm 0.10}$	$40.30_{\pm 1.10}$	$33.80_{\pm 0.93}$
MB Fano	$\mathbf{2.24_{\pm 0.12}}$	$2.49_{\pm 0.19}$	$\mathbf{1.80_{\pm 0.08}}$	$2.25_{\pm 0.14}$	$\mathbf{2.01_{\pm 0.11}}$	$3.67_{\pm 0.13}$	$\mathbf{1.66_{\pm 0.09}}$	$\mathbf{1.89_{\pm 0.06}}$	$\mathbf{14.61_{\pm 0.84}}$	$21.68_{\pm 1.44}$
DPI	$\mathbf{2.24_{\pm 0.17}}$	$2.64_{\pm 0.07}$	$\mathbf{1.73_{\pm 0.07}}$	$2.08_{\pm 0.06}$	$\mathbf{1.98_{\pm 0.09}}$	$4.07_{\pm 0.23}$	$\mathbf{1.64_{\pm 0.07}}$	$\mathbf{1.97_{\pm 0.08}}$	$17.55_{\pm 1.33}$	$\mathbf{17.41_{\pm 0.62}}$

Appendix : Empirical Bernstein Inequality

Hoeffding's Inequality

Let $Z, Z_1, \ldots, Z_n$ be i.i.d. random variables with values in $[0, 1]$ and let $\delta > 0$. Then with probability at least $1 - \delta$ in $(Z_1, \ldots, Z_n)$ we have

$$\mathbb{E}Z - \frac{1}{n}\sum_{i=1}^{n} Z_i \leq \sqrt{\frac{log 1/\delta}{2n}}.$$

Bennett's Inequality

Under the conditions of Hoeffding's inequality, we have with probability at least $1 - \delta$ that

$$\mathbb{E}Z - \frac{1}{n}\sum_{i=1}^{n} Z_i \leq \sqrt{\frac{2\mathbb{V}(Z)\, log 1/\delta}{n}} + \frac{log 1/\delta}{3n}$$

where $\mathbb{V}(Z) = \mathbb{E}(Z - \mathbb{E}Z)$.

Appendix : Empirical Bernstein Inequality

Empirical Bernstein Inequality

Under the conditions of Hoeffding's inequality, we have with probability with at least $1 - \delta$ in the i.i.d. vector $\mathbf{Z} = (Z_1, \ldots, Z_n)$ that

$$\mathbb{E}Z - \frac{1}{n}Z_i \leq \sqrt{\frac{2V_n\left(\mathbf{Z}\right) log2/\delta}{n}} + \frac{7log2/\delta}{3\left(n-1\right)}$$

where $V_n\left(\mathbf{Z}\right) = \frac{1}{n(n-1)}\sum_{1\leq i<j\leq n}\left(Z_i - Z_j\right)^2$

Appendix : Fano's Inequality

Fano's Inequality

Let Z, Y be discrete random variables on $\{1, \ldots, M\}$. Then

$$\mathbb{P}(Z \neq Y) \geq \frac{H(Y|Z) - log2}{logM}$$

Appendix : ConfTr

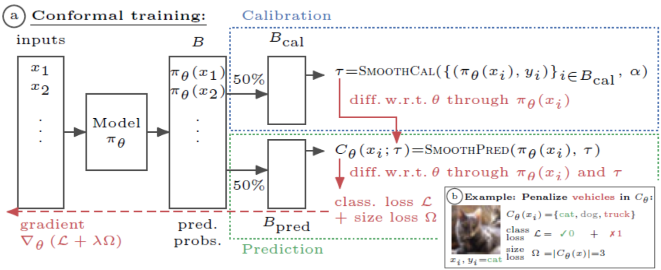
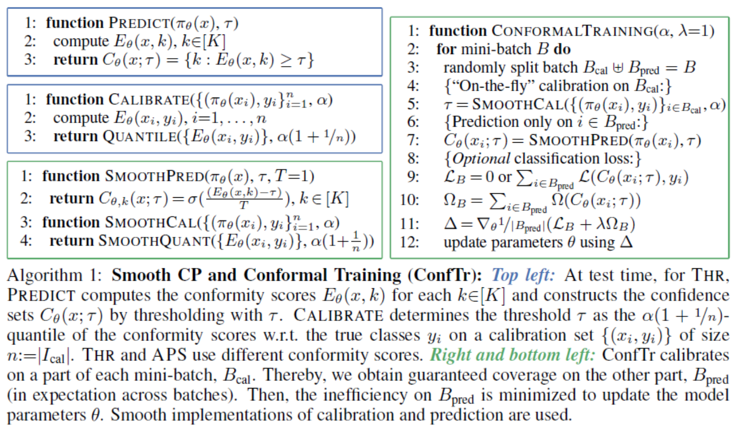

a Conformal training:
Calibration
inputs
B
B_{cal}
x_1
x_2
Model
π_θ
$\pi_\theta(x_1)$
$\pi_\theta(x_2)$
50%
τ=SmoothCal($\{(\pi_\theta(x_i), y_i)\}_{i \in B_{cal}}$, α)
diff. w.r.t. θ through $\pi_\theta(x_i)$
$C_\theta(x_i;\tau)$=SmoothPred($\pi_\theta(x_i)$, τ)
diff. w.r.t. θ through $\pi_\theta(x_i)$ and τ
50%
class. loss $\mathcal{L}$
+ size loss Ω
gradient
$\nabla_\theta(\mathcal{L} + \lambda\Omega)$
pred. probs.
B_{pred}
Prediction
b Example: Penalize vehicles in C_θ:
$C_\theta(x_i)$={cat, dog, truck}
class loss $\mathcal{L}$ = ✓0 + ✗1
size loss Ω =$|C_\theta(x)|$=3
x_i, y_i=cat

Figure 1: **Illustration of *conformal training (ConfTr)*:** We develop differentiable prediction and calibration steps for *conformal prediction (CP)*, SmoothCal and SmoothPred. During training, this allows ConfTr to "simulate" CP on each mini-batch B by calibrating on the first half B_{cal} and predicting confidence sets on the other half B_{pred} (*c.f.* ⓐ). ConfTr can optimize arbitrary losses on the predicted confidence sets, *e.g.*, reducing average confidence set size (*inefficiency*) using a size loss Ω or penalizing specific classes from being included using a classification loss $\mathcal{L}$ (*c.f.* ⓑ). *After* training using our method, *any* existing CP method can be used to obtain a coverage guarantee.

Appendix : ConfTr



```
function PREDICT($\pi_\theta(x), \tau$)
  compute $E_\theta(x, k)$, $k\in[K]$
  return $C_\theta(x;\tau) = \{k : E_\theta(x, k) \geq \tau\}$
```

```
function CALIBRATE($\{(\pi_\theta(x_i), y_i\}_{i=1}^n, \alpha$)
  compute $E_\theta(x_i, y_i)$, $i{=}1, \dots, n$
  return QUANTILE($\{E_\theta(x_i, y_i)\}, \alpha(1 + 1/n)$)
```

```
function SMOOTHPRED($\pi_\theta(x), \tau, T{=}1$)
  return $C_{\theta,k}(x;\tau) = \sigma(\frac{(E_\theta(x,k)-\tau)}{T})$, $k \in [K]$
function SMOOTHCAL($\{(\pi_\theta(x_i), y_i\}_{i=1}^n, \alpha$)
  return SMOOTHQUANT($\{E_\theta(x_i, y_i)\}, \alpha(1+\frac{1}{n})$)
```

```
function CONFORMALTRAINING($\alpha, \lambda{=}1$)
  for mini-batch $B$ do
    randomly split batch $B_{\text{cal}} \uplus B_{\text{pred}} = B$
    {"On-the-fly" calibration on $B_{\text{cal}}$:}
    $\tau =$ SMOOTHCAL($\{(\pi_\theta(x_i), y_i)\}_{i\in B_{\text{cal}}}, \alpha$)
    {Prediction only on $i \in B_{\text{pred}}$:}
    $C_\theta(x_i;\tau) =$ SMOOTHPRED($\pi_\theta(x_i), \tau$)
    {Optional classification loss:}
    $\mathcal{L}_B = 0$ or $\sum_{i\in B_{\text{pred}}} \mathcal{L}(C_\theta(x_i;\tau), y_i)$
    $\Omega_B = \sum_{i\in B_{\text{pred}}} \Omega(C_\theta(x_i;\tau))$
    $\Delta = \nabla_\theta 1/|B_{\text{pred}}|(\mathcal{L}_B + \lambda\Omega_B)$
    update parameters $\theta$ using $\Delta$
```

Algorithm 1: **Smooth CP and Conformal Training (ConfTr):** *Top left:* At test time, for THR, PREDICT computes the conformity scores $E_\theta(x, k)$ for each $k{\in}[K]$ and constructs the confidence sets $C_\theta(x;\tau)$ by thresholding with τ. CALIBRATE determines the threshold τ as the $\alpha(1 + 1/n)$-quantile of the conformity scores w.r.t. the true classes y_i on a calibration set $\{(x_i, y_i)\}$ of size $n{:=}|I_{\text{cal}}|$. THR and APS use different conformity scores. *Right and bottom left:* ConfTr calibrates on a part of each mini-batch, B_{cal}. Thereby, we obtain guaranteed coverage on the other part, B_{pred} (in expectation across batches). Then, the inefficiency on B_{pred} is minimized to update the model parameters θ. Smooth implementations of calibration and prediction are used.

Appendix : THR and APS

- THR (Threshold Conformal Predictor) constructs the cofidence sets by thresholding probabilities $C_\theta\left(x;\tau\right) := \left\{k : \pi_{\theta,k}\left(x\right) \geq \tau\right\}$. τ is computed as the $\alpha\left(1+1/|I_{cal}|\right)$-quantile of the so-called conformity scores $E_\theta\left(x_i, y_i\right) = \pi_{\theta,y_i}\left(x_i\right)$.
- APS (Adpative Prediction Sets) constructs confidence sets based on the ordered probabilities. Specifically, $C_\theta\left(x;\tau\right) := \left\{k : E_\theta\left(x,k\right) \leq \tau\right\}$ with :

$$E_\theta\left(x,k\right) := \pi_{\theta,y^{(1)}}\left(x\right) + \cdots + \pi_{\theta,y^{(k-1)}}\left(x\right) + U\pi_{\theta,y^{(k)}}\left(x\right)$$

where $\pi_{\theta,y^{(1)}}\left(x\right) \geq \cdots \geq \pi_{\theta,y^{(K)}}\left(x\right)$ and U is a uniform random variable in $[0,1]$ to break ties.