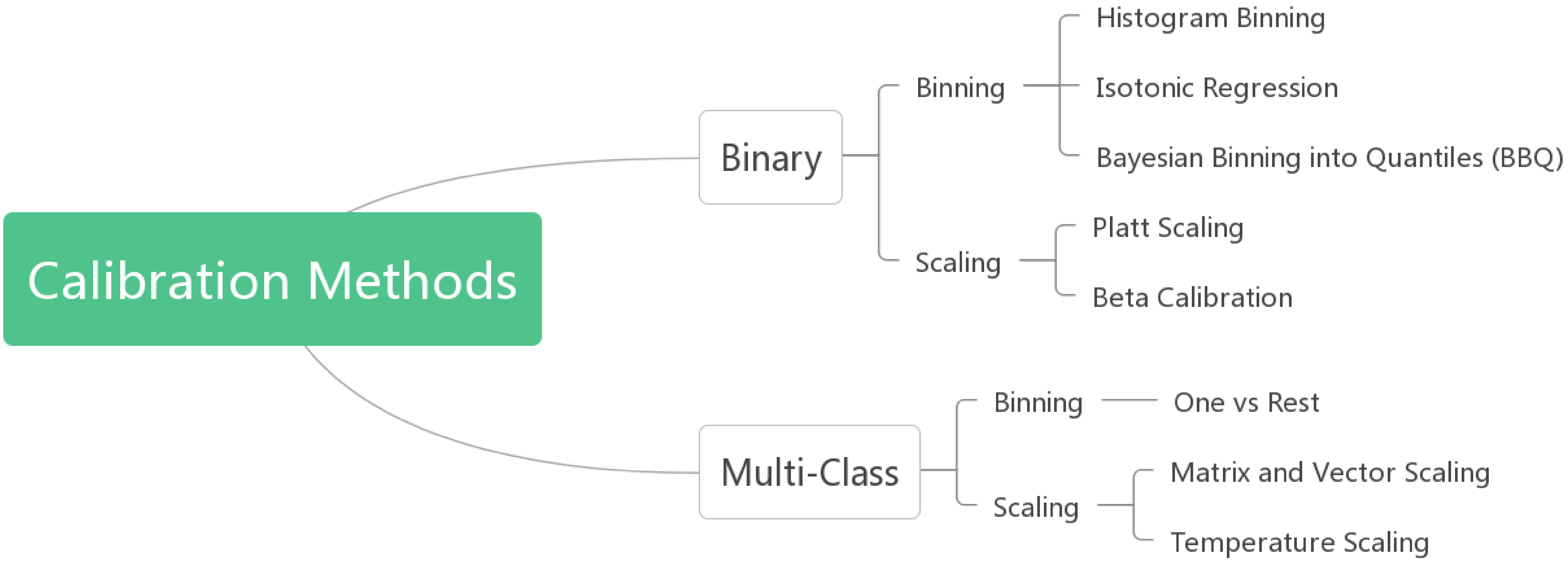
南京航空航天大學
Nanjing University of Aeronautics and Astronautics

ParNeC 模式识别与神经计算研究组
PAttern Recognition and NEural Computing

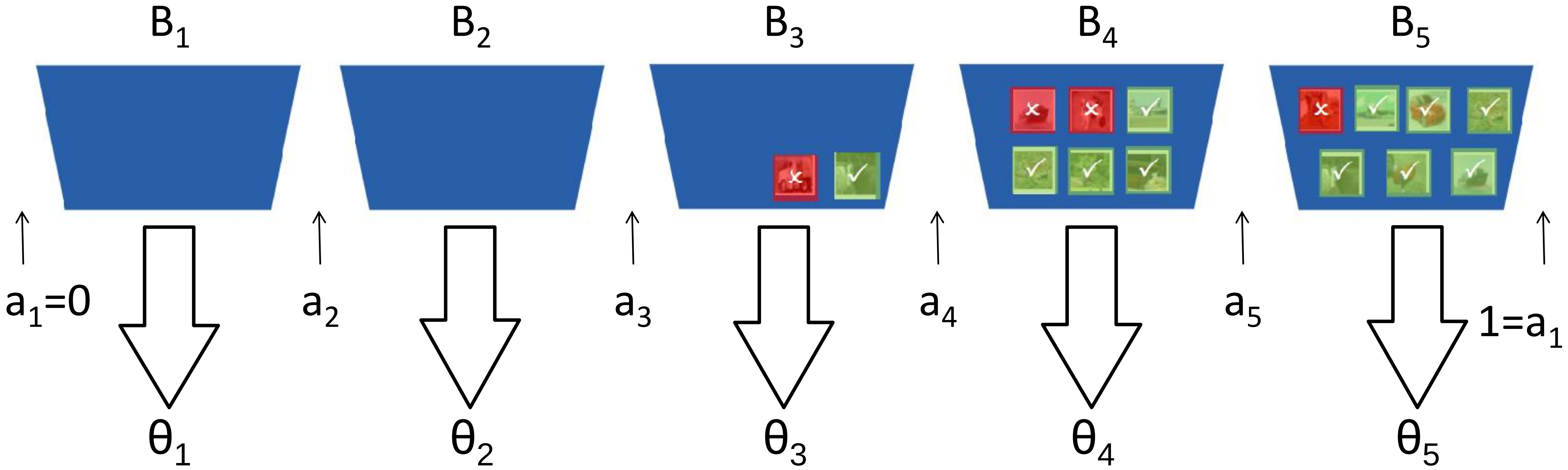
SOTA of Two Technical Routes in Model Calibration: Gaussian Process Calibration & I-Max Binning

Calibration Methods



Binary

Histogram Binning

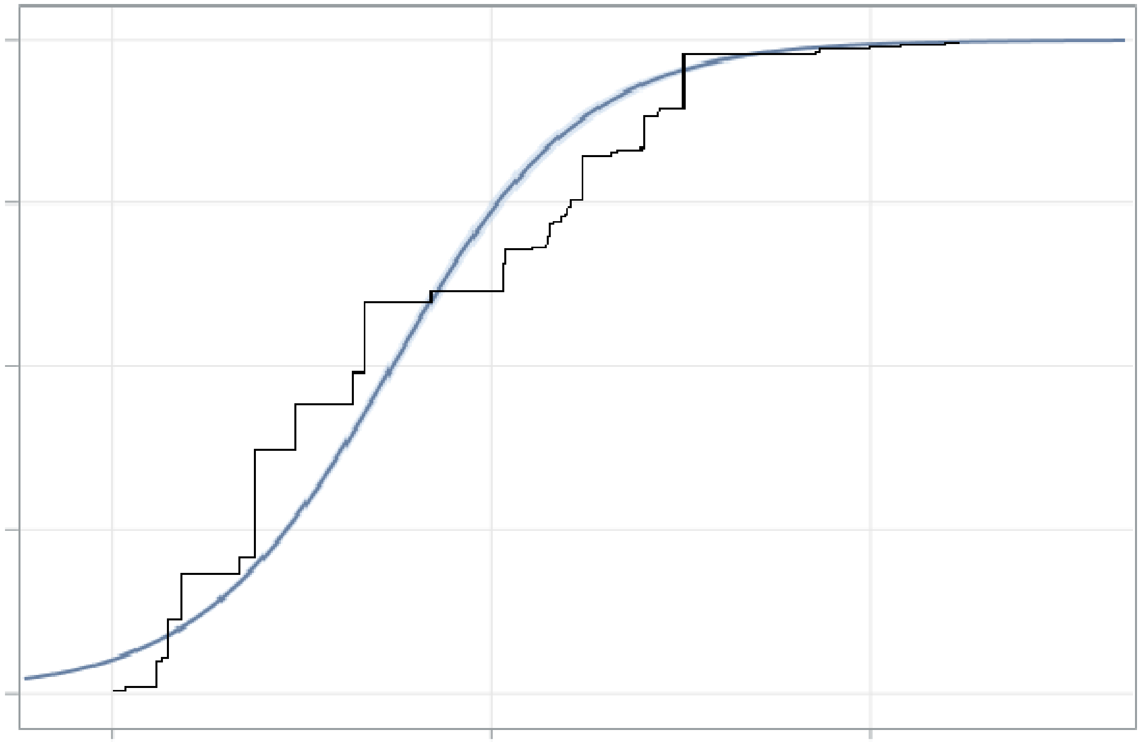


$$\min_{\theta_1,\ldots,\theta_M} \sum_{m=1}^{M}\sum_{i=1}^{n} \mathbf{1}(a_m \le \hat{p}_i < a_{m+1})\,(\theta_m - y_i)^2$$

Given fixed bins boundaries, the solution results in θ_m that correspond to the average number of positive-class samples in bin B_m

Isotonic Regression

-- a strict version of histogram binning where boundaries and predictions are jointly optimized



$$\min_{\substack{M \\ \theta_1,\ldots,\theta_M \\ a_1,\ldots,a_{M+1}}} \sum_{m=1}^{M} \sum_{i=1}^{n} \mathbf{1}(a_m \le \hat{p}_i < a_{m+1}) \left(\theta_m - y_i\right)^2 \quad \text{subject to} \quad \begin{aligned} &0 = a_1 \le a_2 \le \ldots \le a_{M+1} = 1, \\ &\theta_1 \le \theta_2 \le \ldots \le \theta_M . \end{aligned}$$

Bayesian Binning into Quantiles (BBQ)

-- BBQ marginalizes out all possible binning schemes

Settings:

A binning scheme s is a pair $(M, \mathcal{I})$ where M is the number of bins, and $\mathcal{I}$ is a corresponding partitioning of $[0, 1]$ into disjoint intervals $(0 = a_1 \leq a_2 \leq \ldots \leq a_{M+1} = 1)$. The parameters of a binning scheme are $\theta_1, \ldots, \theta_M$.

BBQ considers a space $\mathcal{S}$ of all possible binning schemes for the validation dataset D.

Bayesian Binning into Quantiles (BBQ)

$$\mathbb{P}(\hat{q}_{te} \mid \hat{p}_{te}, D) = \sum_{s \in \mathcal{S}} \mathbb{P}(\hat{q}_{te}, S = s \mid \hat{p}_{te}, D)$$

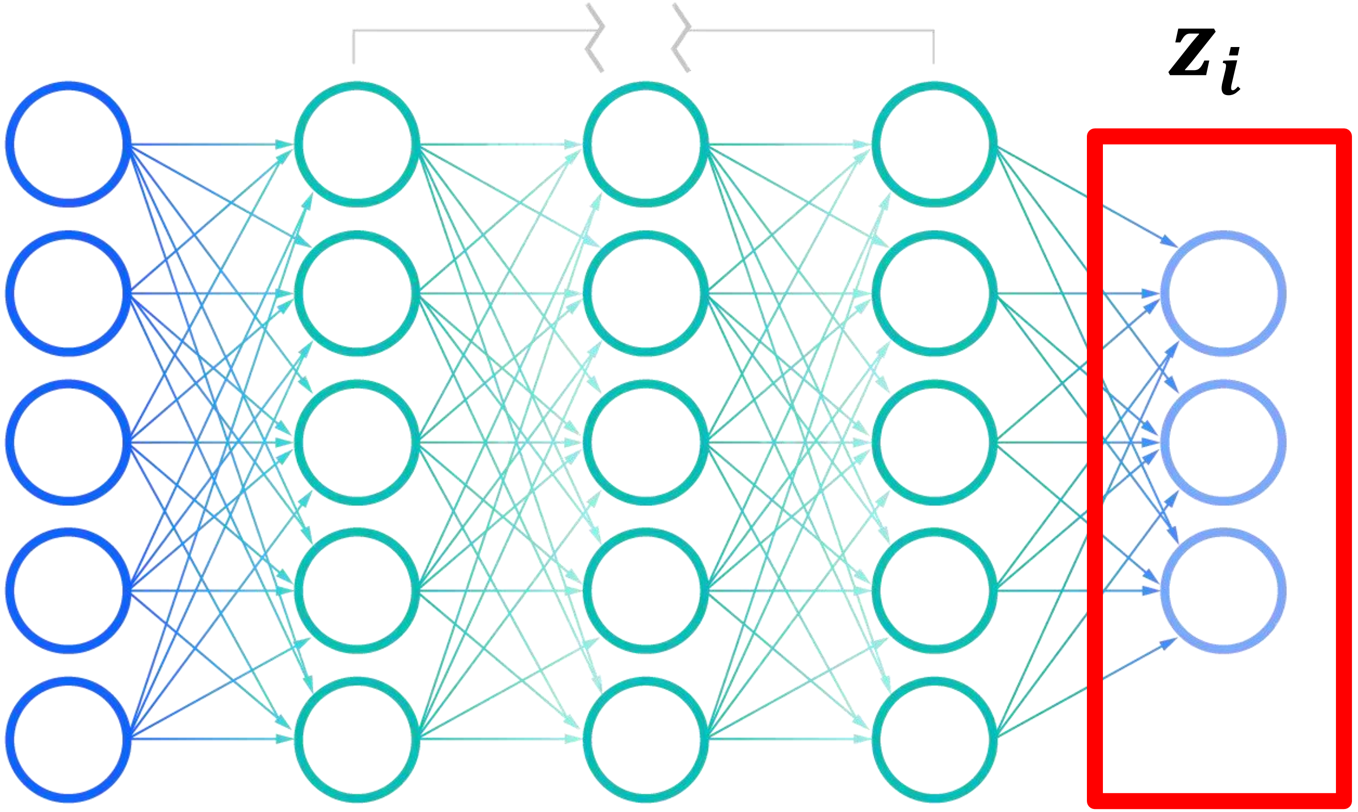
$$= \sum_{s \in \mathcal{S}} \boxed{\mathbb{P}(\hat{q}_{te} \mid \hat{p}_{te}, S=s, D)} \boxed{\mathbb{P}(S=s \mid D)}$$

calibrated probability using binning scheme s

$$\mathbb{P}(S=s \mid D) = \frac{\mathbb{P}(D \mid S=s)}{\sum_{s' \in \mathcal{S}} \mathbb{P}(D \mid S=s')}$$

The parameters $\theta_1, \ldots, \theta_M$ can be viewed as parameters of M independent binomial distributions. Hence, by placing a Beta prior on $\theta_1, \ldots, \theta_M$, we can obtain a closed form expression for the marginal likelihood $\mathbb{P}(D \mid S = s)$. This allows us to compute $\mathbb{P}(\hat{q}_{te} \mid \hat{p}_{te}, D)$ for any test input.

Platt Scaling



$$\hat{q}_i = \sigma(\underline{a} z_i + \underline{b})$$

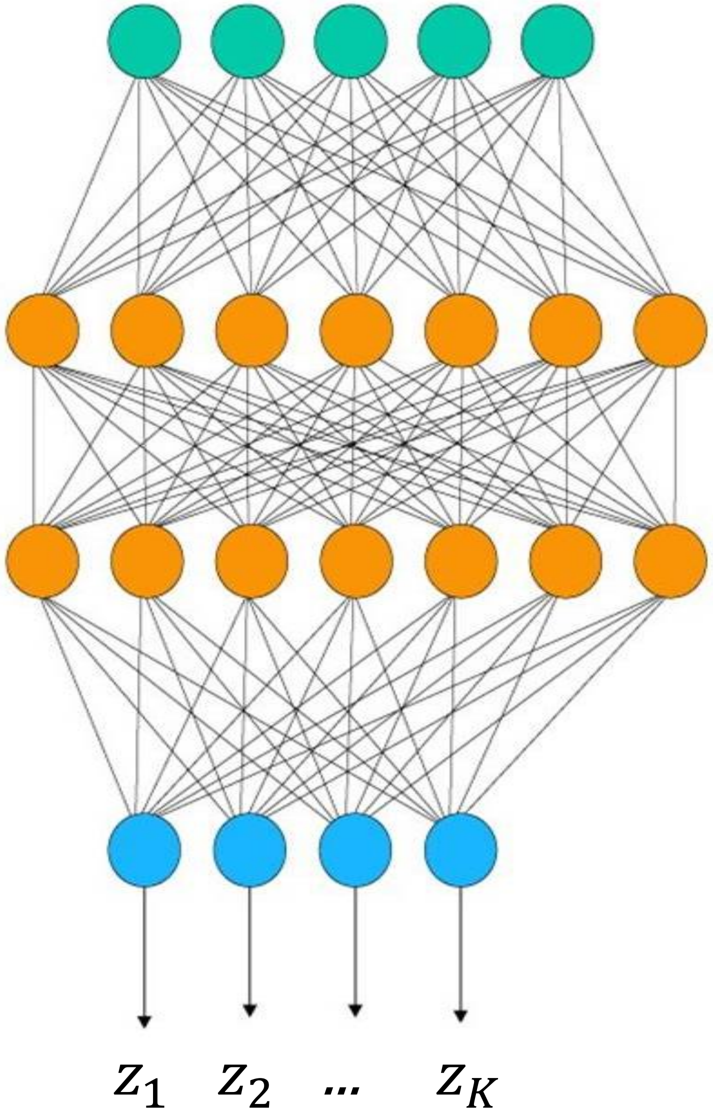
optimized using the NLL loss over the validation set

Beta Calibration

$$\mu_{beta}(s;a,b,c) = \frac{1}{1 + 1\Big/\left(e^{c}\frac{s^{a}}{(1-s)^{b}}\right)}$$

Multi-Class

Extension of Binning Methods (One vs Rest)



We have K classes.

For each class k, we form a binary calibration problem where the label is $\mathbb{I}(y_i = k)$ and the predicted probability is $\sigma_{SM}(z_i)^{(k)}$.

For each instance i, we have an unnormalized probability vector $[\hat{q}_i^{(1)}, \hat{q}_i^{(2)}, \ldots, \hat{q}_i^{(K)}]$.

Then normalize them.

Matrix and Vector Scaling

$$\hat{q}_i = \sigma(az_i + b)$$

$$\Downarrow$$

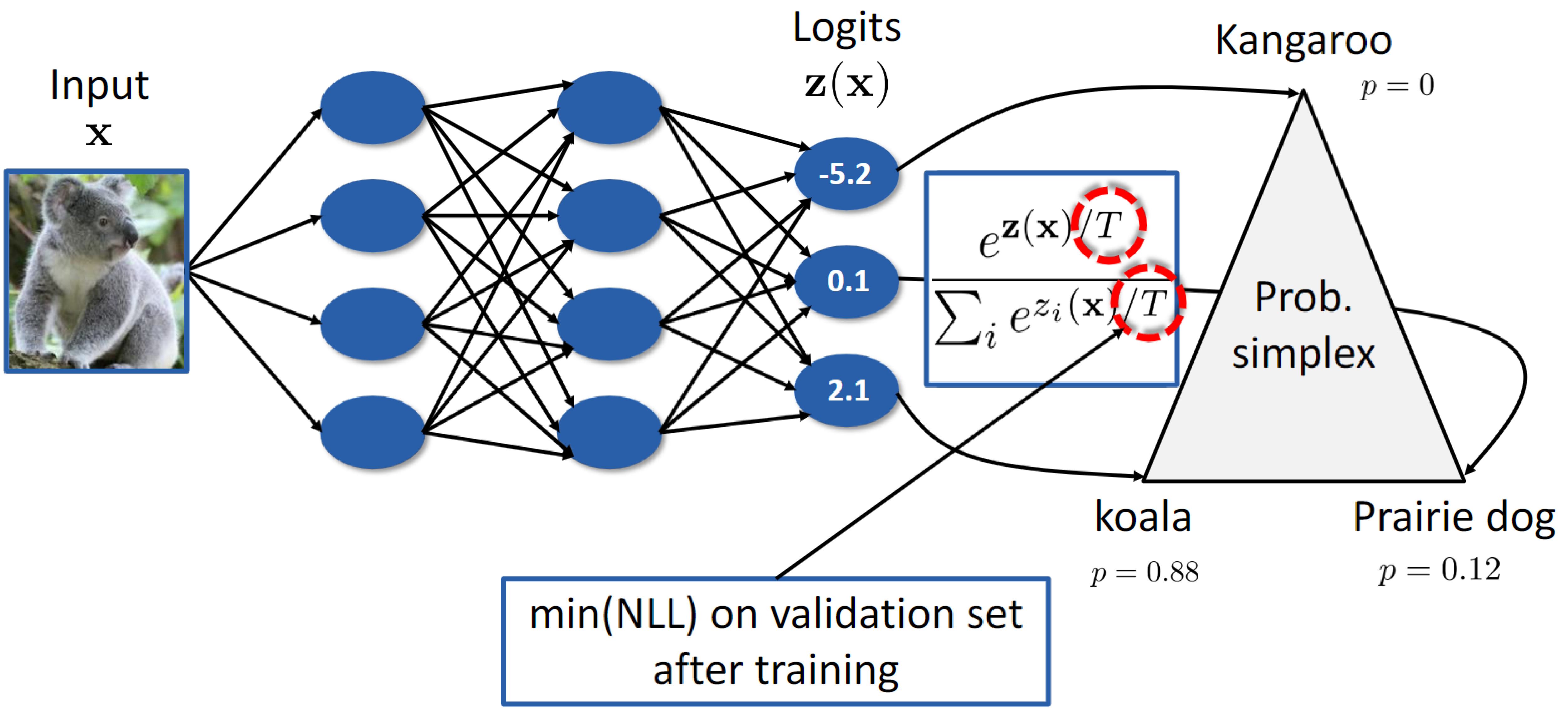
a and b are optimized using the NLL loss over the validation set

$$\hat{q}_i = \max_k \sigma_{\text{SM}}(\mathbf{W}\mathbf{z}_i + \mathbf{b})^{(k)}$$

$$\hat{y}_i' = \operatorname*{argmax}_k (\mathbf{W}\mathbf{z}_i + \mathbf{b})^{(k)}$$

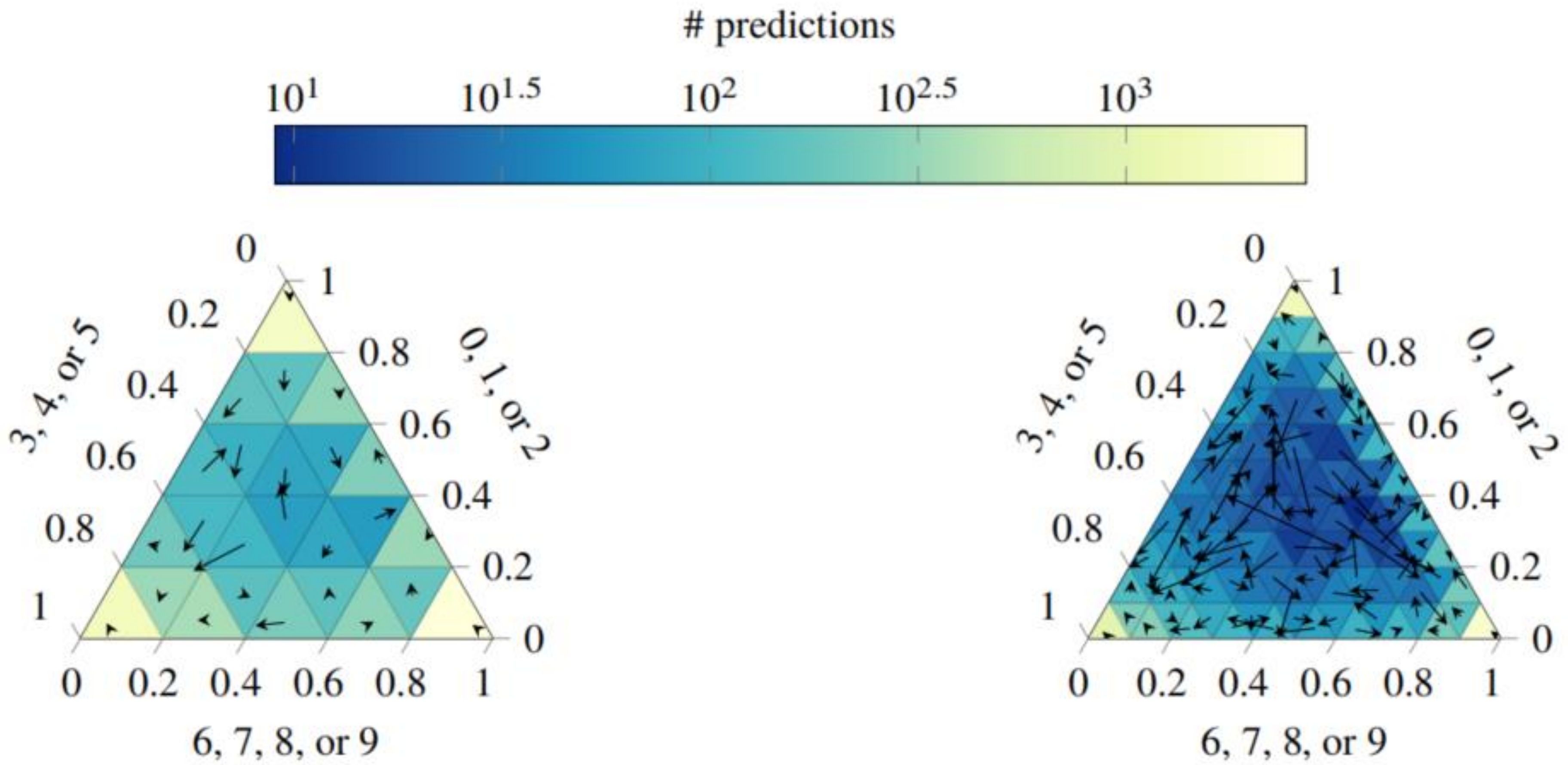
W and b are optimized using the NLL loss over the validation set

Temperature Scaling



An Interesting Visualization Method

An Interesting Visualization Method



LeNet on CIFAR-10

智周万物
道济天下

Non-Parametric Calibration for Classification

Jonathan Wenger　Hedvig Kjellström　Rudolph Triebel

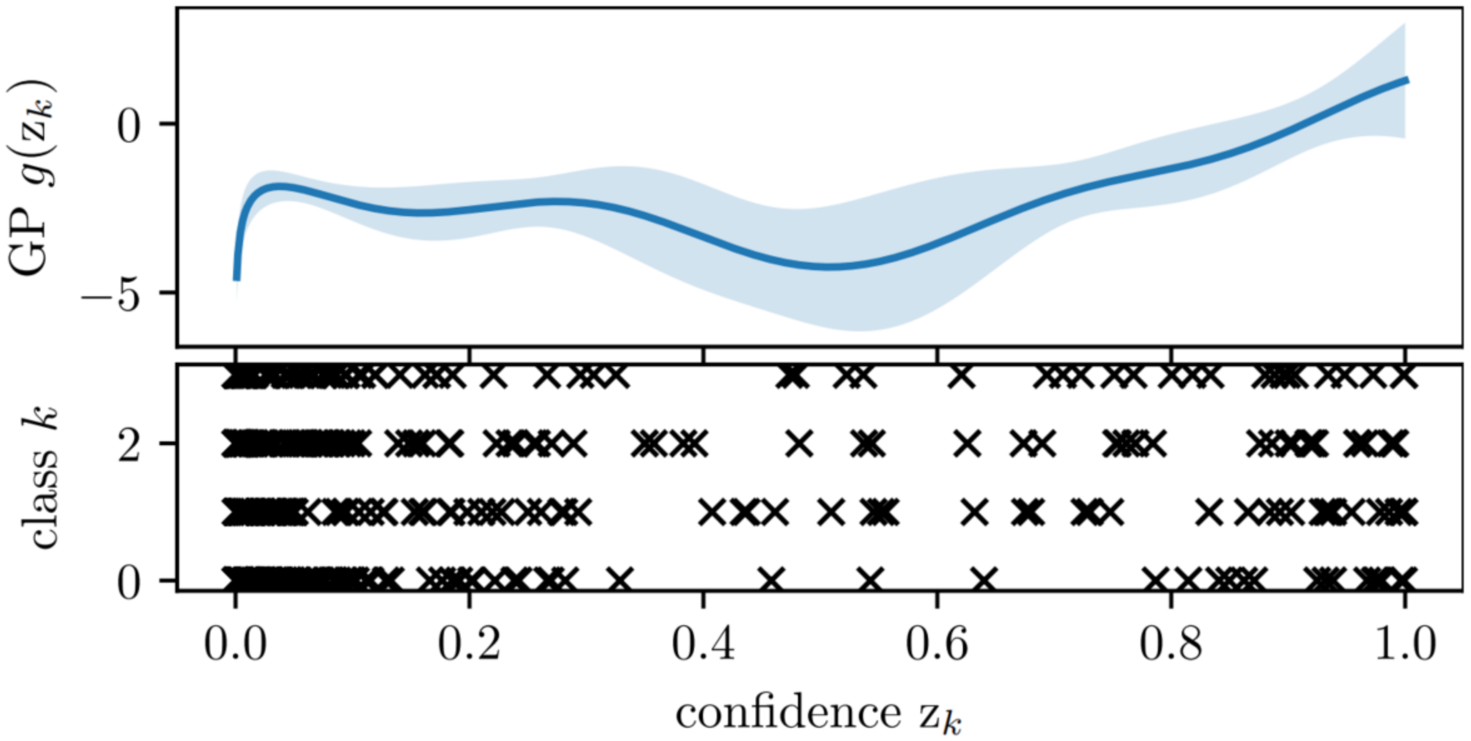
AISTATS 2020

Gaussian Process

Definition Assume a one-dimensional Gaussian process prior over the latent function $g : \mathbb{R} \to \mathbb{R}$, i.e.

$$g \sim \mathcal{GP}\left(\mu(\cdot), k(\cdot\,, \cdot \mid \boldsymbol{\theta})\right)$$

with mean function μ, kernel k and kernel parameters $\boldsymbol{\theta}$



The Joint Distribution of the data & latent variables

model output:

$$v(\mathbf{z})_k = \sigma(g(\mathrm{z}_1), \ldots, g(\mathrm{z}_K))_k = \frac{\exp(g(\mathrm{z}_k))}{\sum_{j=1}^{K} \exp(g(\mathrm{z}_j))}$$

categorical likelihood:

$$\mathrm{Cat}(\mathrm{y} \mid v(\mathbf{z})) = \prod_{k=1}^{K} \sigma(g(\mathrm{z}_1), \ldots, g(\mathrm{z}_K))_k^{[\mathrm{y}=k]}$$

The joint distribution of the data $(\boldsymbol{z}_n, y_n)$ and latent variables $\boldsymbol{g}$:

$$p(\mathbf{y}, \mathbf{g}) = p(\mathbf{y} \mid \mathbf{g}) p(\mathbf{g}) = \prod_{n=1}^{N} p(\mathrm{y}_n \mid \mathbf{g}_n) p(\mathbf{g}) = \prod_{n=1}^{N} \mathrm{Cat}(\mathrm{y}_n \mid \sigma(\mathbf{g}_n)) \mathcal{N}(\mathbf{g} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}_{\mathbf{g}})$$

where $\mathbf{y} \in \{1, \ldots, K\}^N$, $\mathbf{g} = (\mathbf{g}_1, \mathbf{g}_2, \ldots, \mathbf{g}_N)^\top \in \mathbb{R}^{NK}$

and $\mathbf{g}_n = (g(\mathrm{z}_{n1}), \ldots, g(\mathrm{z}_{nK}))^\top \in \mathbb{R}^K$

Using Variational Inference

In order to reduce the computational complexity $\mathcal{O}((NK)^3)$, we define M inducing inputs $\mathbf{w} \in \mathbb{R}^M$ and inducing variables $\mathbf{u} \in \mathbb{R}^M$.

$$p(\mathbf{g}, \mathbf{u}) = \mathcal{N}\left(\begin{bmatrix}\mathbf{g} \\ \mathbf{u}\end{bmatrix} \middle| \begin{bmatrix}\boldsymbol{\mu}_{\mathbf{g}} \\ \boldsymbol{\mu}_{\mathbf{u}}\end{bmatrix}, \begin{bmatrix}\boldsymbol{\Sigma}_{\mathbf{g}} & \boldsymbol{\Sigma}_{\mathbf{g},\mathbf{u}} \\ \boldsymbol{\Sigma}_{\mathbf{g},\mathbf{u}}^\top & \boldsymbol{\Sigma}_{\mathbf{u}}\end{bmatrix}\right)$$

$q(\mathbf{u}) = \mathcal{N}(\mathbf{u} \mid \boldsymbol{m}, \boldsymbol{S})$: a variational approximation to the posterior $p(\mathbf{u} \mid \boldsymbol{y})$

$$\boxed{p(\mathbf{g}, \mathbf{u} \mid \mathbf{y}) = p(\mathbf{g} \mid \mathbf{u})\, p(\mathbf{u} \mid \mathbf{y})}$$

Using Variational Inference

$$
\begin{aligned}
\ln p(\mathbf{y}) &\geq \mathrm{ELBO}(q(\mathbf{u})) \\
&= \mathbb{E}_{q(\mathbf{u})}\left[\ln p(\mathbf{y} \mid \mathbf{u})\right] - \mathrm{KL}\left[q(\mathbf{u}) \| p(\mathbf{u})\right] \\
&\geq \mathbb{E}_{q(\mathbf{u})}\left[\mathbb{E}_{p(\mathbf{g}|\mathbf{u})}\left[\ln p(\mathbf{y} \mid \mathbf{g})\right]\right] \\
&\qquad - \mathrm{KL}\left[q(\mathbf{u}) \| p(\mathbf{u})\right] \\
&= \mathbb{E}_{q(\mathbf{g})}\left[\ln p(\mathbf{y} \mid \mathbf{g})\right] - \mathrm{KL}\left[q(\mathbf{u}) \| p(\mathbf{u})\right]
\end{aligned}
$$

(Considering $\mathrm{KL}(q(\boldsymbol{u})||p(\boldsymbol{u}|\boldsymbol{y}))$)

Let $q(\mathbf{u}) = \mathcal{N}(\mathbf{u} \mid \boldsymbol{m}, \boldsymbol{S})$ and $\boldsymbol{A} := \boldsymbol{\Sigma}_{\mathbf{g},\mathbf{u}} \boldsymbol{\Sigma}_{\mathbf{u}}^{-1}$, then

$$
\begin{aligned}
q(\mathbf{g}) &:= \int \underbrace{p(\mathbf{g} \mid \mathbf{u}) q(\mathbf{u})}_{q(\mathbf{g},\mathbf{u})} \, d\mathbf{u} \\
&= \mathcal{N}(\mathbf{g} \mid \boldsymbol{\mu}_{\mathbf{g}} + \boldsymbol{A}(\boldsymbol{m} - \boldsymbol{\mu}_u), \boldsymbol{\Sigma}_{\mathbf{g}} + \boldsymbol{A}(\boldsymbol{S} - \boldsymbol{\Sigma}_{\mathbf{u}})\boldsymbol{A}^\top)
\end{aligned}
$$

$q(\boldsymbol{g}) := \int p(\boldsymbol{g}|\boldsymbol{u})q(\boldsymbol{u})d\boldsymbol{u}$ is Gaussian, its K-dimensional marginals $q(\boldsymbol{g}_n) = \mathcal{N}(\boldsymbol{g}_n|\boldsymbol{\varphi}_n, \boldsymbol{C}_n)$

Using Variational Inference

南京航空航天大学 Nanjing University of Aeronautics and Astronautics

(Considering $\mathrm{KL}(q(\boldsymbol{u})||p(\boldsymbol{u}|\boldsymbol{y}))$)

$$\begin{aligned}
\ln p(\mathbf{y}) &\geq \mathrm{ELBO}(q(\mathbf{u})) \\
&= \mathbb{E}_{q(\mathbf{u})}\left[\ln p(\mathbf{y} \mid \mathbf{u})\right] - \mathrm{KL}\left[q(\mathbf{u}) \| p(\mathbf{u})\right] \\
&\geq \mathbb{E}_{q(\mathbf{u})}\left[\mathbb{E}_{p(\mathbf{g}|\mathbf{u})}\left[\ln p(\mathbf{y} \mid \mathbf{g})\right]\right] \\
&\qquad - \mathrm{KL}\left[q(\mathbf{u}) \| p(\mathbf{u})\right] \\
&= \mathbb{E}_{q(\mathbf{g})}\left[\ln p(\mathbf{y} \mid \mathbf{g})\right] - \mathrm{KL}\left[q(\mathbf{u}) \| p(\mathbf{u})\right] \\
&= \sum_{n=1}^{N} \mathbb{E}_{q(\mathbf{g}_n)}\left[\ln p(\mathrm{y}_n \mid \mathbf{g}_n)\right] - \mathrm{KL}\left[q(\mathbf{u}) \| p(\mathbf{u})\right]
\end{aligned}$$

$$\begin{aligned}
\mathbb{E}_{q(\mathbf{g}_n)}\left[\ln p(\mathrm{y}_n \mid \mathbf{g}_n)\right] \approx \ln p(\mathrm{y}_n \mid \boldsymbol{\phi}_n) \\
+ \frac{1}{2}\left(\sigma(\boldsymbol{\phi}_n)^\top \boldsymbol{C}_n \sigma(\boldsymbol{\phi}_n) - \mathrm{diag}(\boldsymbol{C}_n)^\top \sigma(\boldsymbol{\phi}_n)\right)
\end{aligned}$$

GP Calibration

$$p(\mathbf{g}, \mathbf{u} \mid \mathbf{y}) \approx p(\mathbf{g} \mid \mathbf{u})q(\mathbf{u})$$

$$\begin{aligned} p(\mathbf{g}_* \mid \mathbf{y}) &= \int p(\mathbf{g}_* \mid \mathbf{g}, \mathbf{u}) p(\mathbf{g}, \mathbf{u} \mid \mathbf{y}) \, d\mathbf{g} \, d\mathbf{u} \\ &\approx \int p(\mathbf{g}_* \mid \mathbf{u}) q(\mathbf{u}) \, d\mathbf{u} \end{aligned}$$

Computational Complexity

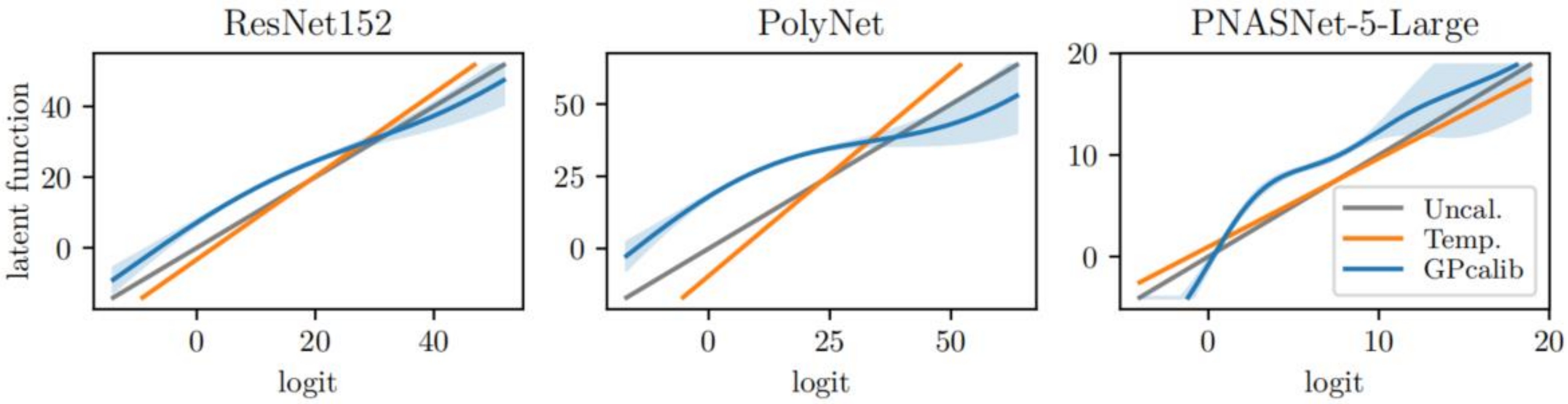
南京航空航天大学 Nanjing University of Aeronautics and Astronautics

Method	Optim. obj.	Calibration
Temp. scal.	$\mathcal{O}(NK)$	$\mathcal{O}(K)$
GPcalib		
diag. cov.	$\mathcal{O}(NK + M^3)$	$\mathcal{O}(K(M^2 + Q))$
full cov.	$\mathcal{O}(NK^2 + M^3)$	$\mathcal{O}(K^2(M^2 + Q))$
mean appr.	$\mathcal{O}(NK^a + M^3)$	$\mathcal{O}(K^a M^2)$

Experiments

Data Set	Model	Uncal.	one-vs-all Platt	one-vs-all Isotonic	one-vs-all Beta	one-vs-all BBQ	Temp.	GPcalib
MNIST	AdaBoost	.6121	.2267	.1319	.2222	.1384	.1567	**.0414**
MNIST	XGBoost	.0740	.0449	**.0176**	**.0184**	.0207	.0222	**.0180**
MNIST	Mondrian Forest	.2163	.0357	.0282	.0383	.0762	**.0208**	**.0213**
MNIST	Random Forest	.1178	.0273	.0207	.0259	.1233	**.0121**	.0148
MNIST	1 layer NN	.0262	**.0126**	**.0140**	.0168	.0186	.0195	.0239
CIFAR-100	AlexNet	.2751	.0720	.1232	.0784	.0478	**.0365**	**.0369**
CIFAR-100	WideResNet	.0664	.0838	.0661	.0539	.0384	.0444	**.0283**
CIFAR-100	ResNeXt-29 (8x64)	.0495	.0882	.0599	.0492	.0392	.0424	**.0251**
CIFAR-100	ResNeXt-29 (16x64)	.0527	.0900	.0620	.0520	.0365	.0465	**.0266**
CIFAR-100	DenseNet-BC-190	.0717	.0801	.0665	.0543	.0376	.0377	**.0237**
ImageNet	AlexNet	**.0353**	.1132	.2937	.2290	.1307	**.0342**	**.0357**
ImageNet	VGG19	.0377	.0965	.2810	.2416	.1617	**.0342**	**.0364**
ImageNet	ResNet-50	.0441	.0875	.2724	.2250	.1635	**.0341**	**.0335**
ImageNet	ResNet-152	.0545	.0879	.2761	.2201	.1675	.0323	**.0283**
ImageNet	DenseNet-121	.0380	.0949	.2682	.2297	.1512	**.0329**	.0357
ImageNet	DenseNet-201	.0410	.0898	.2706	.2189	.1614	**.0324**	.0367
ImageNet	InceptionV4	.0318	.0865	.2900	.1653	.1593	.0462	**.0269**
ImageNet	SE-ResNeXt-50	.0440	.0889	.2684	.1789	.1990	.0482	**.0279**
ImageNet	SE-ResNeXt-101	.0574	.0853	.2844	.1631	.1496	.0415	**.0250**
ImageNet	PolyNet	.0823	.0806	.2590	.2006	.1787	.0369	**.0283**
ImageNet	SENet-154	.0612	.0809	.3003	.1582	.1502	.0497	**.0309**
ImageNet	PNASNet-5-Large	.0702	.0796	.3063	.1430	.1355	.0486	**.0270**
ImageNet	NASNet-A-Large	.0530	.0826	.3265	.1437	.1268	.0516	**.0255**

Experiments



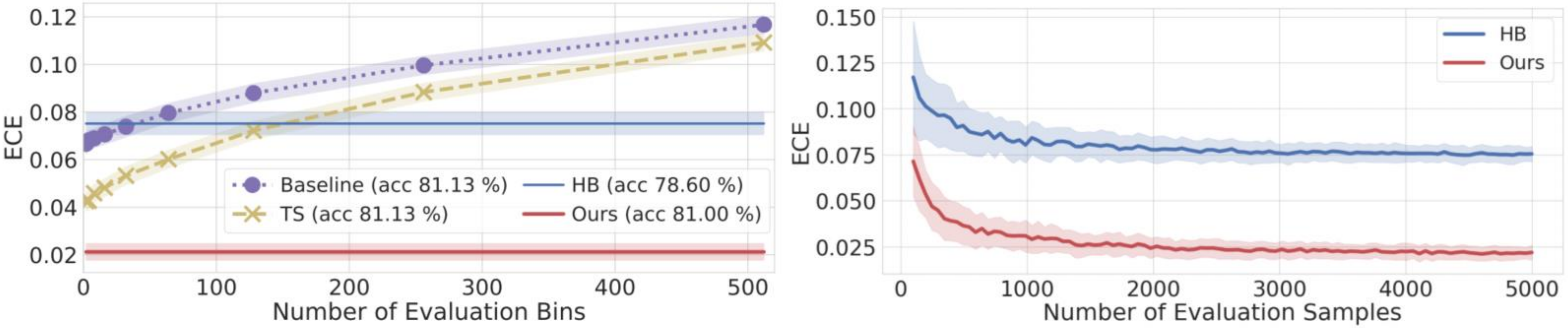
The plot shows latent functions of temperature scaling and GPcalib from a single CV run of our experiments on ImageNet. For PolyNet and PNASNet GPcalib shows a significant decrease in ECE1 in Table 1, corresponding to a higher degree of non-linearity in the latent GP.

Multi-Class Uncertainty Calibration via Mutual Information Maximization-based Binning

Kanil Patel[1,2], William Beluch[1], Bin Yang[2], Michael Pfeiffer[1], Dan Zhang[1]
[1] Bosch Center for Artificial Intelligence, Renningen, Germany
[2] Institute of Signal Processing and System Theory, University of Stuttgart, Stuttgart, Germany

ICLR 2021 Poster

HB's Advantage



(a) Top-1 prediction ECE (5k evaluation samples) (b) ECE converging curve (based on 10^2 bootstraps)

Figure 1: (a) Temperature scaling (TS), equally sized-histogram binning (HB), and our proposal, i.e., sCW I-Max binning are compared for post-hoc calibrating a CIFAR100 (WRN) classifier. (b) Binning offers a reliable ECE measure as the number of evaluation samples increases.

HB's ECE estimate is constant and unaffected by the number of evaluation bins.

Problem Setup

input: $x \in \mathcal{X}$ belongs to one of K classes

ground truth lables: $\mathbf{y} = [y_1, y_2, \ldots, y_K] \in \{0,1\}^K$

Let $f : \mathcal{X} \mapsto [0,1]^K$,

output: $\mathbf{q} = [q_1, \ldots, q_K] \in [0,1]^K$

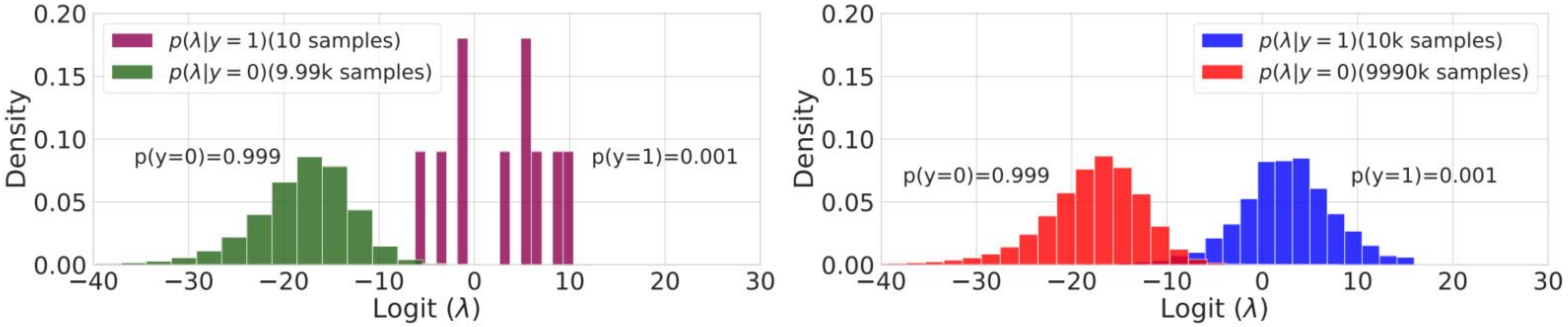
class-wise ECE: ${}_{\mathrm{cw}}\mathrm{ECE}(h \circ f) = \frac{1}{K} \sum_{k=1}^{K} E_{\mathbf{q}=f(\mathbf{x})} \left\{ \left| p(y_k = 1 | h(\mathbf{q})) - h_k(\mathbf{q}) \right| \right\}$

top-1 ECE: $E\left[\left|p(y_{k=\arg\max_k h_k(\mathbf{q})} = 1 | h(\mathbf{q})) - \max_k h_k(\mathbf{q})\right|\right]$

make q unbounded: $\lambda_k \triangleq \log q_k - \log(1 - q_k)$

quantizer Q: $\lambda \in \mathbb{R} \rightarrow m \in \{1, \ldots, M\}$ if $\lambda \in \mathcal{I}_m = [g_{m-1}, g_m)$,
where M is the total number of bin intervals, $g_{m-1} < g_m, g_0 = -\infty, g_M = \infty$.
Any logit binned to $\mathcal{I}_m$ will be reproduced to the same bin representative r_m.

shared class-wise(sCW) vs CW



(a) train set $S_{k=394}$ for CW binning.

(b) train. set S for *shared*-CW binning.

Figure A1: Histogram of ImageNet (InceptionResNetv2) logits for (a) CW and (b) sCW training. By means of the set merging strategy to handle the two-class imbalance $1:999$, S has K=1000 times more class-1 samples than S_k with the same 10k calibration samples from C.

shared class-wise(sCW) vs CW

南京航空航天大学
Nanjing University of Aeronautics and Astronautics

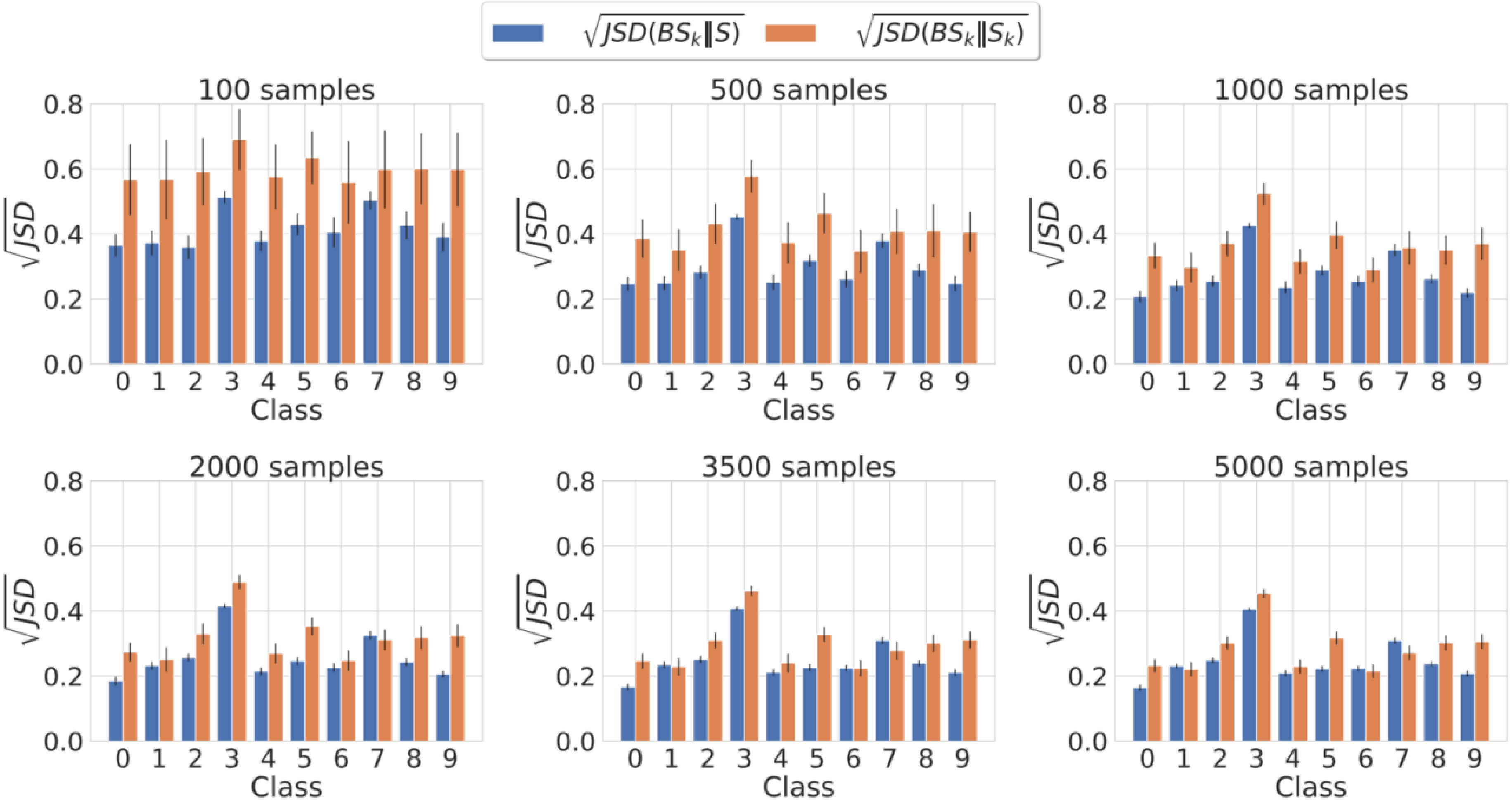


Figure A2: Empirical approximation error of S vs. S_k, where Jensen-Shannon divergence (JSD) is used to measure the difference between the empirical distributions underlying the training sets for class-wise bin optimization. Overall, the merged set S is a more sample efficient choice over S_k.

Mutual Information(MI) Maximization

We propose bin optimization via maximizing the MI between the quantized logits $Q(\lambda)$ and and the label y:

$$\{g_m^*\} = \arg \max_{Q:\{g_m\}} I(y; m = Q(\lambda)) \overset{(a)}{=} \arg \max_{Q:\{g_m\}} H(m) - H(m|y)$$

$$P(m) = \int_{g_{m-1}}^{g_m} p(\lambda) \mathrm{d}\lambda \qquad P(m|y) = \int_{g_{m-1}}^{g_m} p(\lambda|y) \mathrm{d}\lambda$$

Mutual Information(MI) Maximization

Theorem 1. *The MI maximization problem given in ($\boxed{2}$) is equivalent to*

$$\max_{Q:\{g_m\}} I(y; m = Q(\lambda)) \equiv \min_{\{g_m,\phi_m\}} \mathcal{L}(\{g_m, \phi_m\})$$

where the loss $\mathcal{L}(\{g_m, \phi_m\})$ *is defined as*

$$\mathcal{L}(\{g_m, \phi_m\}) \triangleq \sum_{m=0}^{M-1} \int_{g_m}^{g_{m+1}} p(\lambda) \sum_{y'\in\{0,1\}} P(y = y'|\lambda) \log \frac{P(y = y')}{\sigma\left[(2y' - 1)\phi_m\right]} \mathrm{d}\lambda$$

$$I(y; Q(\lambda)) + \mathcal{L}(\{g_m, \phi_m\}) = \sum_{i=0}^{M-1} \int_{g_m}^{g_{m+1}} p(\lambda) \sum_{y'\in\{0,1\}} P(y = y'|\lambda) \mathrm{d}\lambda \log \frac{P(y = y'|m)}{P_\sigma(y = y'; \phi_m)}$$

$$\overset{(a)}{=} \sum_{i=0}^{M-1} P(m) \left[\sum_{y'\in\{0,1\}} P(y = y'|m) \log \frac{P(y = y'|m)}{P_\sigma(y = y'; \phi_m)} \right]$$

$$\overset{(b)}{=} \sum_{i=0}^{M-1} P(m) \mathrm{KLD}\left[P(y = y'|m) \| P_\sigma(y = y'; \phi_m)\right] \geq 0$$

I-Max Binning

南京航空航天大学 Nanjing University of Aeronautics and Astronautics

$$g_m = \log\left\{\frac{\log\left[\frac{1+e^{\phi_m}}{1+e^{\phi_{m-1}}}\right]}{\log\left[\frac{1+e^{-\phi_{m-1}}}{1+e^{-\phi_m}}\right]}\right\},\quad \phi_m = \log\left\{\frac{\int_{g_m}^{g_{m+1}}\sigma(\lambda)p(\lambda)\mathrm{d}\lambda}{\int_{g_m}^{g_{m+1}}\sigma(-\lambda)p(\lambda)\mathrm{d}\lambda}\right\} \approx \log\left\{\frac{\sum_{\lambda_n\in\mathcal{S}_m}\sigma(\lambda_n)}{\sum_{\lambda_n\in\mathcal{S}_m}\sigma(-\lambda_n)}\right\}$$

Algorithm 1: I-Max Binning Calibration

Input: Number of bins M, logits $\{\lambda_n\}_1^N$ and binary labels $\{y_n\}_1^N$

Result: bin edges $\{g_m\}_0^M$ ($g_0 = -\infty$ and $g_M = \infty$) and bin representations $\{\phi_m\}_0^{M-1}$

Initialization: $\{\phi_m\} \leftarrow$ Kmeans++($\{\lambda_n\}_1^N$, M) (see $\boxed{\text{A3.4}}$) ;

for $iteration = 1, 2, \ldots, 200$ **do**

  for $m = 1, 2, \ldots, M-1$ **do**

$$g_m \leftarrow \log\left\{\frac{\log\left[\frac{1+e^{\phi_m}}{1+e^{\phi_{m-1}}}\right]}{\log\left[\frac{1+e^{-\phi_{m-1}}}{1+e^{-\phi_m}}\right]}\right\};$$

  end

  for $m = 0, 2, \ldots, M-1$ **do**

$$\mathcal{S}_m \triangleq \{\lambda_n\} \cap [g_m, g_{m+1}) ;$$

$$\phi_m \leftarrow \log\left\{\frac{\sum_{\lambda_n\in\mathcal{S}_m}\sigma(\lambda_n)}{\sum_{\lambda_n\in\mathcal{S}_m}\sigma(-\lambda_n)}\right\};$$

  end

end

Compare with classic HB methods

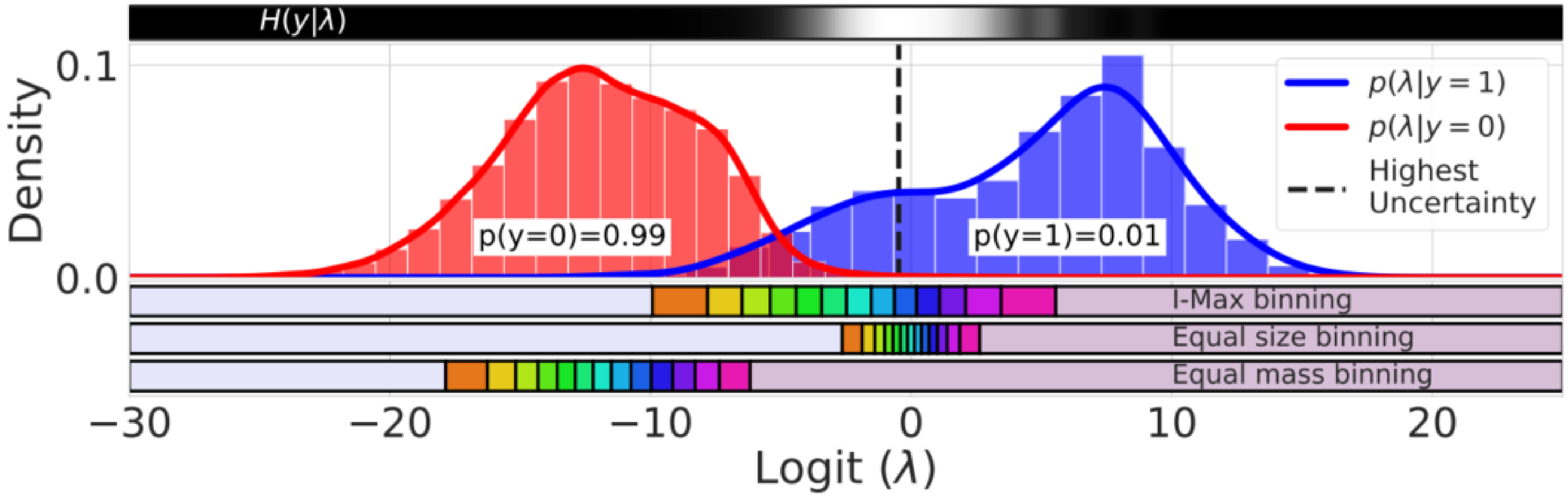
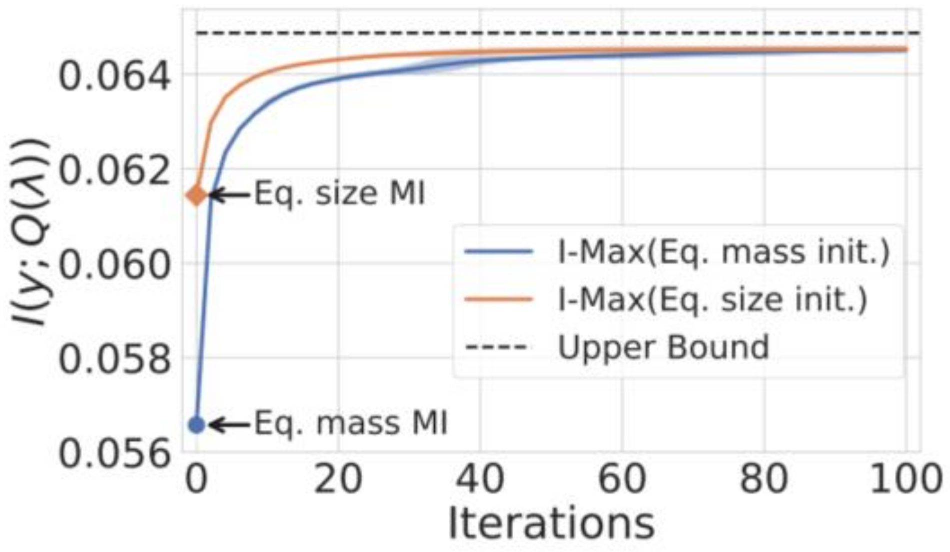


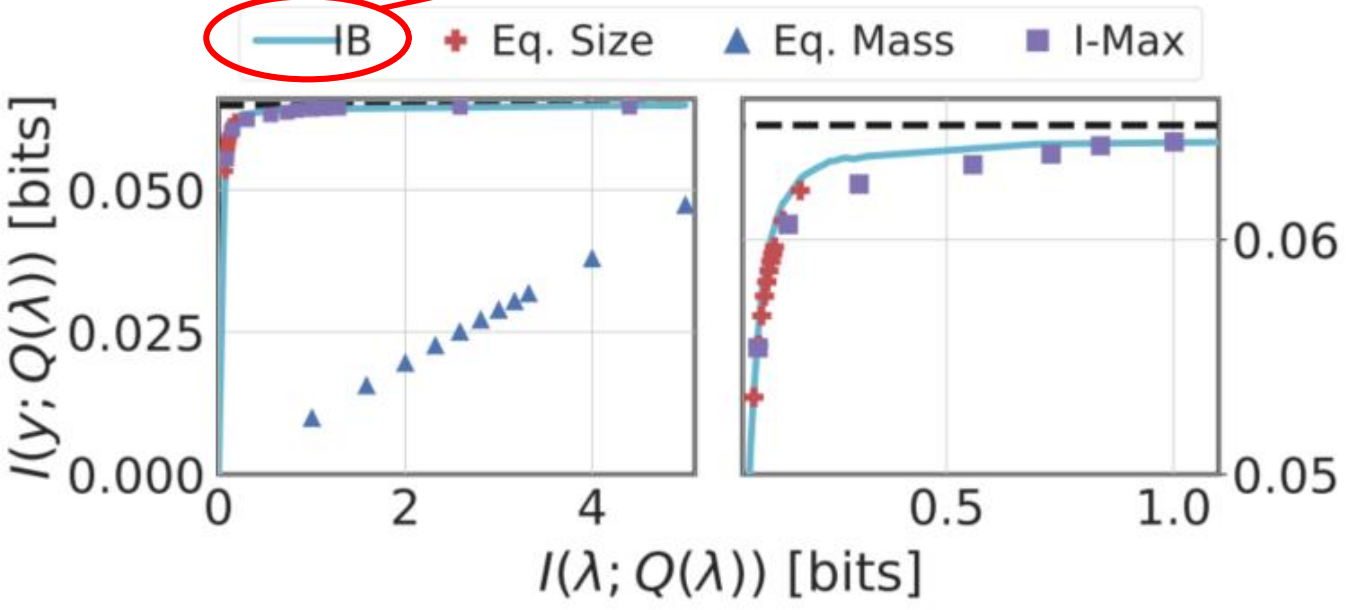
Figure 2: Histogram and KDE of CIFAR100 (WRN) logits in S constructed from 1k calibration samples. The bin edges of Eq. mass binning are located at the high mass region, mainly covering class-0 due to the imbalanced two class ratio $1:99$. Both Eq. size and I-Max binning cover the high uncertainty region, but here only I-Max yields reasonable bin widths ensuring enough mass per bin. Note, Eq. size binning uniformly partitions the interval $[0,1]$ in the probability domain. The observed dense and symmetric bin location around zero is the outcome of probability-to-logit translation.

Compare with classic HB methods

Information Bottleneck limit



(a) Convergence behavior



(b) Label-information vs. compression rate

Experiments

Table A7: Tab. 1 Extension: ImageNet - ResNet152

Binn.	sCW(?)	size	$\text{Acc}_{\text{top1}} \uparrow$	$\text{Acc}_{\text{top5}} \uparrow$	$_{\text{CW}}\text{ECE}_{\frac{1}{K}} \downarrow$	$_{\text{top1}}\text{ECE} \downarrow$	NLL
Baseline	✗	-	**78.33** $\pm$ 0.17	**94.00** $\pm$ 0.14	0.0500 $\pm$ 0.0004	0.0512 $\pm$ 0.0018	0.8760 $\pm$ 0.0133
Eq. Mass	✗	25k	17.45 $\pm$ 0.10	44.87 $\pm$ 0.37	0.0017 $\pm$ 0.0000	0.1555 $\pm$ 0.0010	2.9526 $\pm$ 0.0168
Eq. Mass	✓	1k	16.25 $\pm$ 0.54	45.53 $\pm$ 0.81	0.0064 $\pm$ 0.0004	0.1476 $\pm$ 0.0054	2.9471 $\pm$ 0.0556
Eq. Size	✗	25k	75.50 $\pm$ 0.28	88.85 $\pm$ 0.19	0.1223 $\pm$ 0.0008	0.0604 $\pm$ 0.0017	1.6012 $\pm$ 0.0252
Eq. Size	✓	1k	78.24 $\pm$ 0.16	88.81 $\pm$ 0.19	0.1480 $\pm$ 0.0015	0.0286 $\pm$ 0.0053	1.3308 $\pm$ 0.0178
I-Max	✗	25k	78.24 $\pm$ 0.16	93.91 $\pm$ 0.17	0.0334 $\pm$ 0.0005	0.0521 $\pm$ 0.0015	0.8842 $\pm$ 0.0135
I-Max	✓	1k	78.19 $\pm$ 0.21	93.82 $\pm$ 0.17	**0.0295** $\pm$ 0.0030	**0.0196** $\pm$ 0.0049	**0.8638** $\pm$ 0.0135

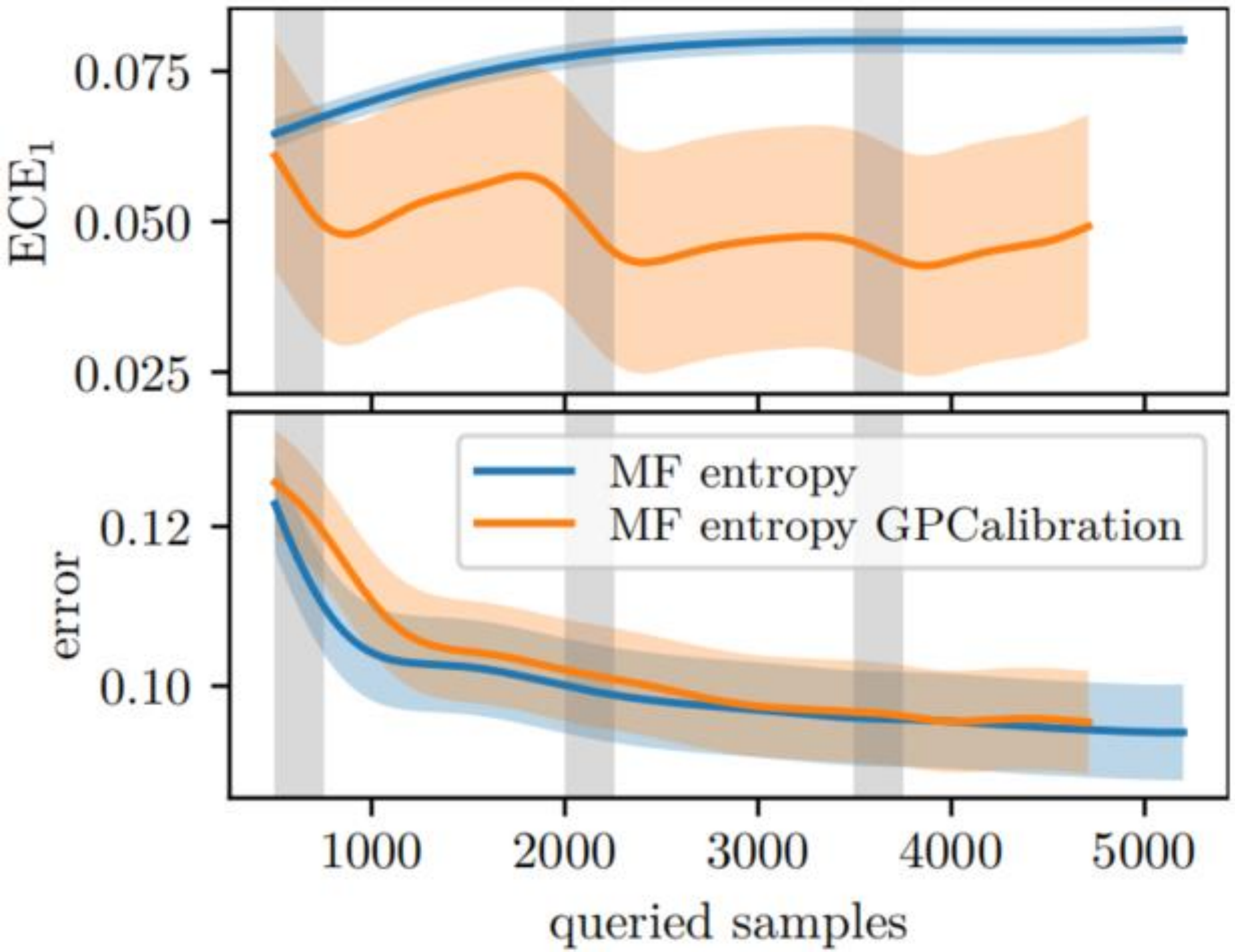
Experiments

ImageNet - InceptionResnetV2

Calibrator	$\mathrm{Acc}_{\mathrm{top1}} \uparrow$	$\mathrm{Acc}_{\mathrm{top5}} \uparrow$	$_{\mathrm{cw}}\mathrm{ECE}_{\frac{1}{K}} \downarrow$	$_{\mathrm{top1}}\mathrm{ECE} \downarrow$	NLL	Brier
Baseline	80.33 ± 0.15	95.10 ± 0.15	0.0486 ± 0.0003	0.0357 ± 0.0009	0.8406 ± 0.0095	0.1115 ± 0.0007
			25k Calibration Samples			
BBQ (Naeini et al., 2015)	53.89 ± 0.30	88.63 ± 0.22	0.0287 ± 0.0009	0.2689 ± 0.0033	1.7104 ± 0.0370	0.3273 ± 0.0016
Beta Kull et al. (2017)	80.47 ± 0.14	94.84 ± 0.15	0.0706 ± 0.0003	0.0346 ± 0.0022	0.9038 ± 0.0270	0.1174 ± 0.0010
Isotonic Reg. Zadrozny & Elkan (2002)	80.08 ± 0.19	93.46 ± 0.20	0.0644 ± 0.0014	0.0468 ± 0.0020	1.8375 ± 0.0587	0.1203 ± 0.0012
Platt Platt (1999)	80.48 ± 0.14	95.18 ± 0.12	0.0597 ± 0.0007	0.0775 ± 0.0015	0.8083 ± 0.0106	0.1205 ± 0.0010
Vec Scal. Kull et al. (2019)	80.53 ± 0.19	95.18 ± 0.16	0.0494 ± 0.0002	0.0300 ± 0.0010	0.8269 ± 0.0097	0.1106 ± 0.0007
Mtx Scal. Kull et al. (2019)	$\mathbf{80.78} \pm 0.18$	$\mathbf{95.38} \pm 0.15$	0.0508 ± 0.0003	0.0282 ± 0.0014	0.8042 ± 0.0100	0.1090 ± 0.0006
BWS Ji et al. (2019)	80.33 ± 0.16	95.10 ± 0.16	0.0561 ± 0.0008	0.044 ± 0.0019	0.8273 ± 0.0105	0.1129 ± 0.0009
ETS-MnM Zhang et al. (2020)	80.33 ± 0.16	95.10 ± 0.16	0.0479 ± 0.0004	0.0358 ± 0.0009	0.8426 ± 0.0097	0.1115 ± 0.0008
			1k Calibration Samples			
TS Guo et al. (2017)	80.33 ± 0.16	95.10 ± 0.16	0.0559 ± 0.0015	0.0439 ± 0.0022	0.8293 ± 0.0107	0.1134 ± 0.0010
GP Wenger et al. (2020)	80.33 ± 0.15	95.11 ± 0.15	0.0485 ± 0.0035	0.0186 ± 0.0034	$\mathbf{0.7556} \pm 0.0118$	$\mathbf{0.1069} \pm 0.0007$
Eq. Mass	5.02 ± 0.13	26.75 ± 0.37	0.0022 ± 0.0001	0.0353 ± 0.0012	3.5272 ± 0.0142	0.0489 ± 0.0012
Eq. Size	80.14 ± 0.23	88.99 ± 0.12	0.1525 ± 0.0023	0.0279 ± 0.0043	1.2671 ± 0.0130	0.1115 ± 0.0011
I-Max	80.20 ± 0.18	94.86 ± 0.17	0.0302 ± 0.0041	0.0200 ± 0.0033	0.7860 ± 0.0208	0.1116 ± 0.0008
Eq. Mass w. TS	5.02 ± 0.13	26.87 ± 0.43	0.0023 ± 0.0001	0.0357 ± 0.0012	3.5454 ± 0.0222	0.0490 ± 0.0012
Eq. Mass w. GP	5.02 ± 0.13	26.87 ± 0.43	0.0022 ± 0.0001	0.0353 ± 0.0012	3.4778 ± 0.0217	0.0489 ± 0.0012
Eq. Size w. TS	80.26 ± 0.18	88.99 ± 0.12	0.1470 ± 0.0007	0.0391 ± 0.0038	1.2721 ± 0.0116	0.1136 ± 0.0012
Eq. Size w. GP	80.26 ± 0.18	88.99 ± 0.12	0.1508 ± 0.0021	0.0140 ± 0.0056	1.2661 ± 0.0121	0.1105 ± 0.0008
I-Max w. TS	80.20 ± 0.18	94.87 ± 0.19	0.0354 ± 0.0124	0.0402 ± 0.0019	0.8339 ± 0.0108	0.1142 ± 0.0009
I-Max w. GP	80.20 ± 0.18	94.87 ± 0.19	$\mathbf{0.0300} \pm 0.0041$	$\mathbf{0.0121} \pm 0.0048$	0.7787 ± 0.0102	0.1111 ± 0.0006

Model Calibration + Active Learning?

Experiments



Mondrian Forests trained online on labels obtained via an entropy query strategy on the KITTI dataset.

The calibrated forest queries about 10% less labels, while reaching comparable accuracy.

Thanks for Listening