Paper Reading

Shuo Wang

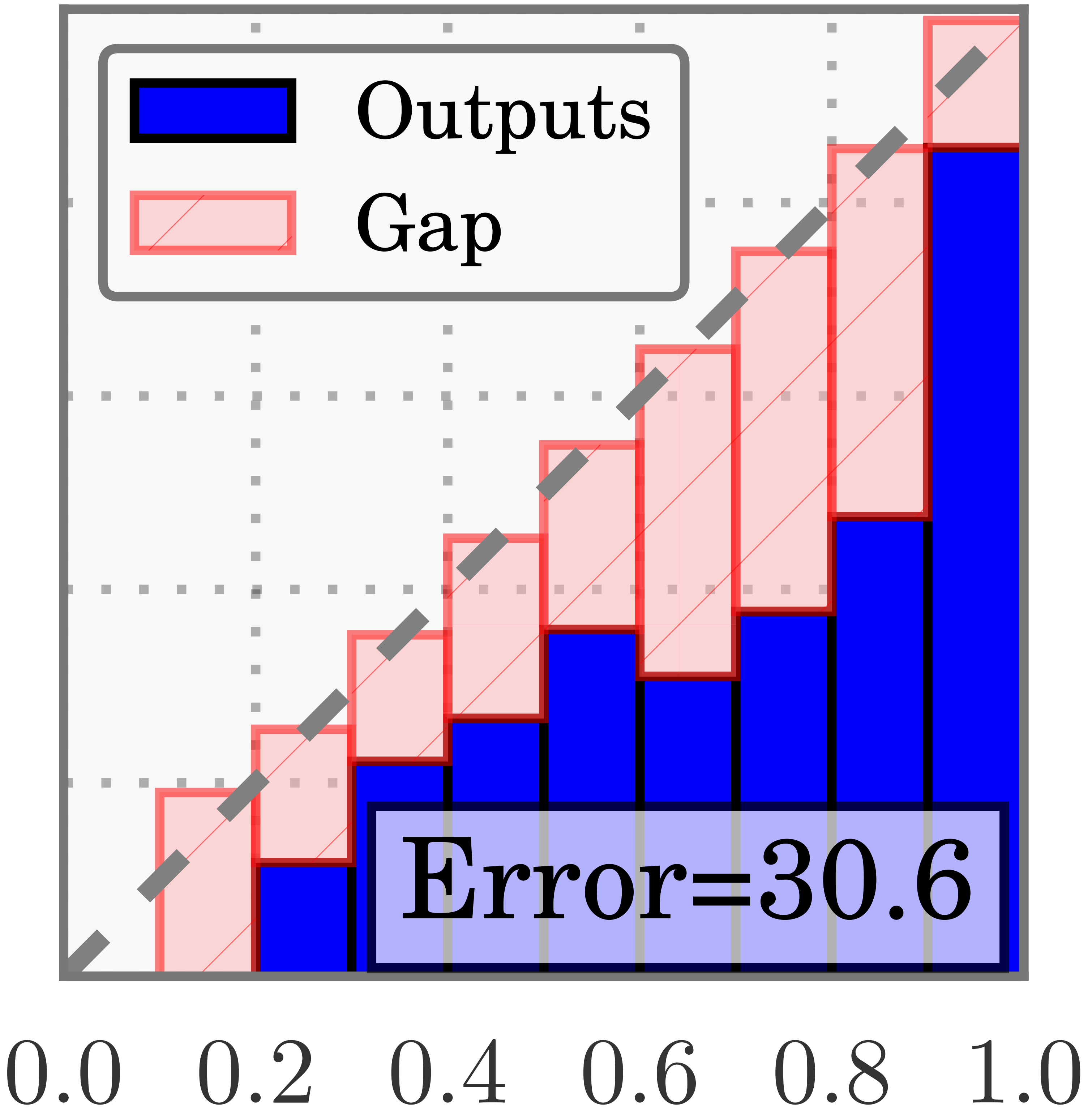
2020-07-02

Overview

- Title: Individual Calibration with Randomized Forecasting
- Authors: Shengjia Zhao, Tengyu Ma, Stefano Ermon
- Affiliation: Computer Science Department, Stanford University

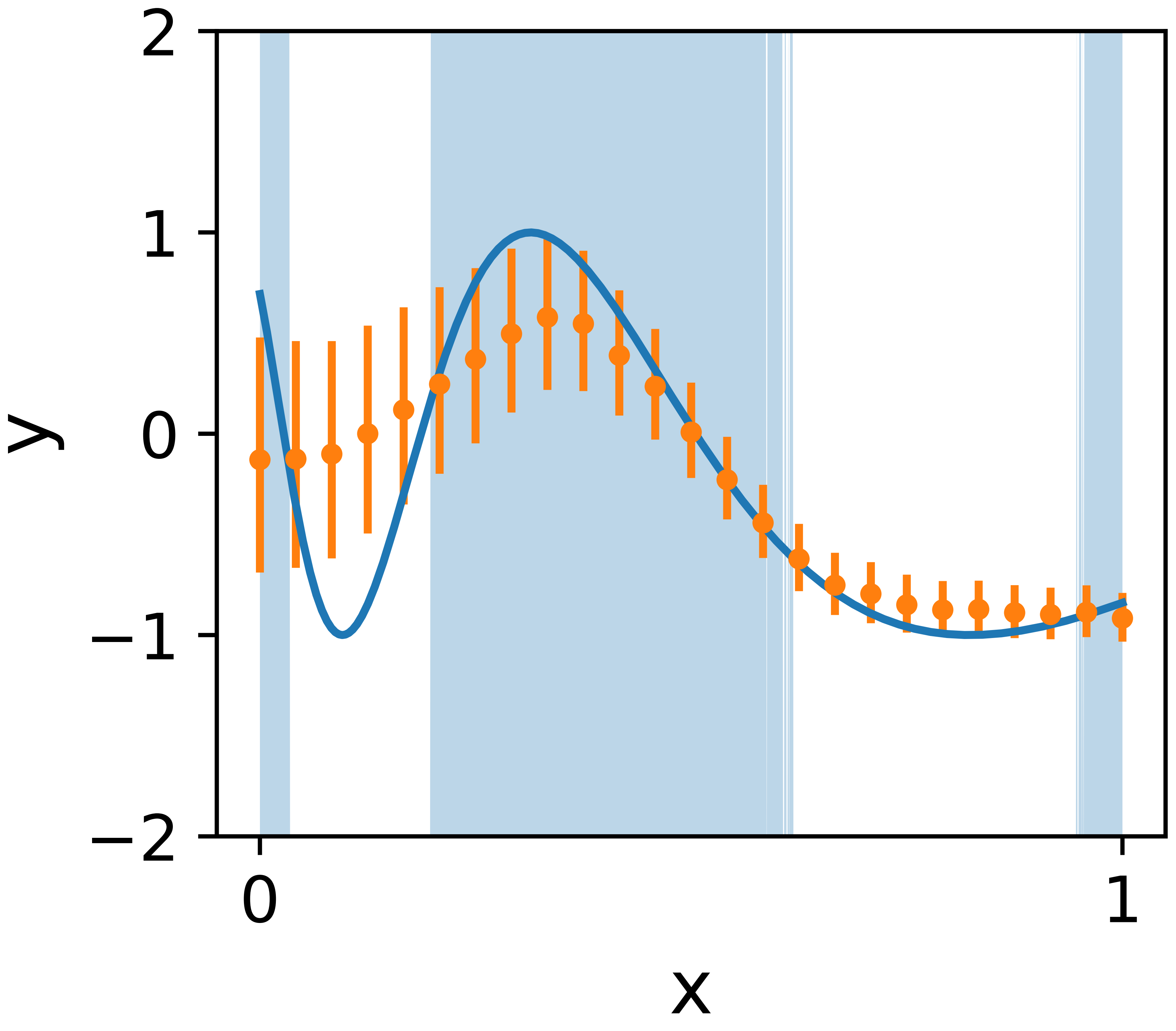
What is Calibration

- Classification
- 对每个输入模型做出预测
- 将概率同属一个区间的预测收集起来，并计算这个集合的平均准确率与平均概率
- 要求平均概率与平均准确率尽可能接近



What is Calibration

- Regression
- 对每个输入模型在输出区间上预测一个概率分布
- 设定一个置信值，根据置信值，确定一个预测的范围
- 统计所有的样本中，ground-truth落在预测范围内的频率
- 要求上述频率和事先确定的置信值尽可能接近



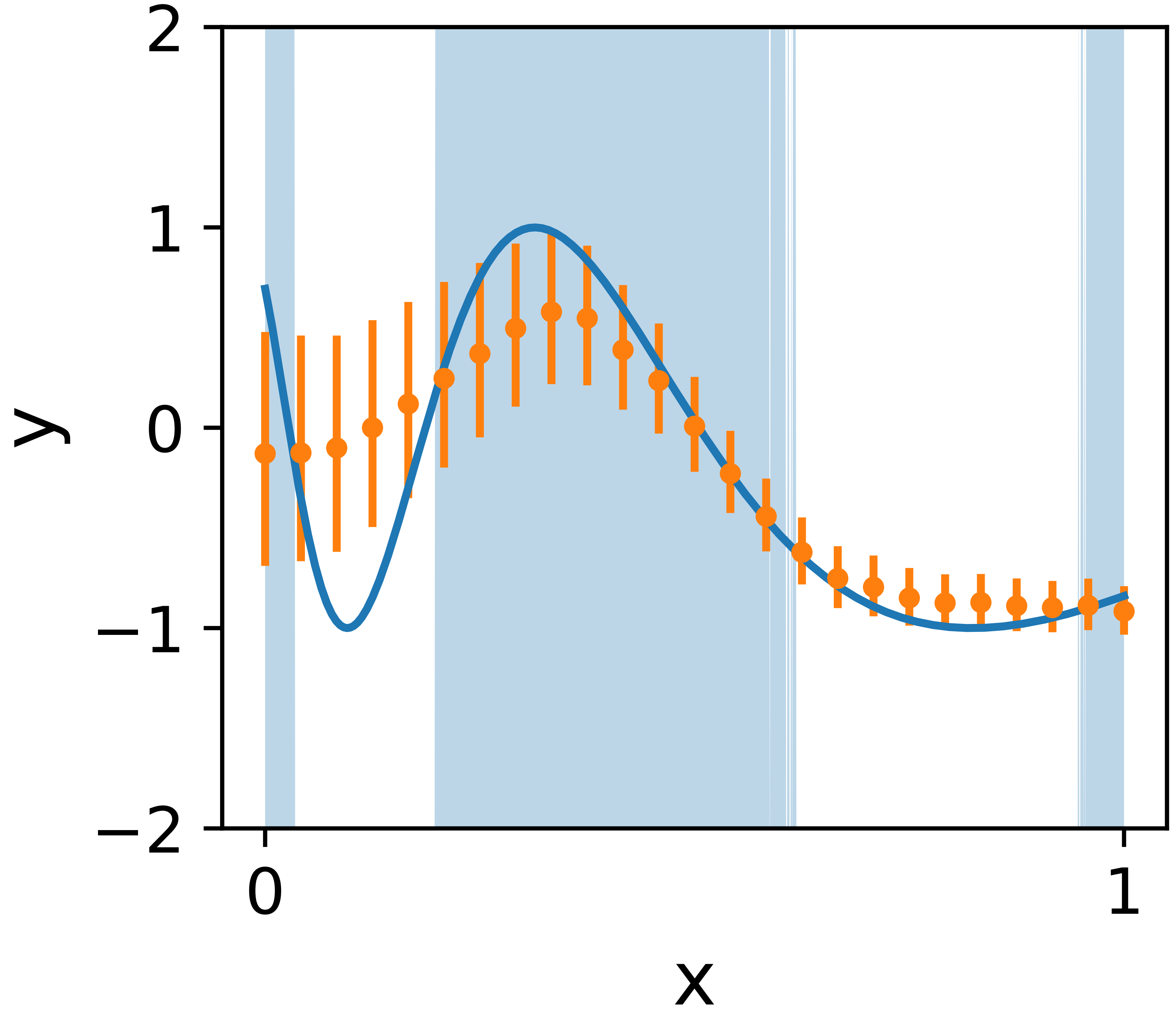
What is Calibration

- Classification
- 对每个输入模型做出预测
- 将概率同属一个区间的预测收集起来，并计算这个集合的平均准确率与平均概率
- 要求平均概率与平均准确率尽可能接近

- Regression
- 对每个输入模型在输出区间上预测一个概率分布
- 设定一个置信值，根据置信值，确定一个预测的范围
- 统计所有的样本中，ground-truth落在预测范围内的频率
- 要求上述频率和事先确定的置信值尽可能接近

The Motivation of Individual Calibration

- Average Calibration的问题：一个满足average calibration的模型，有可能对称地在一个subgroup上over estimate，在另一个group上under estimate (比如不同性别).



Preliminary: Forecaster and Calibration

$\mathbf{X}$ is a random variable on $\mathcal{X}$, if $\mathcal{S} \subset \mathcal{X}$ is a measurable set and $\Pr[\mathbf{X} \in \mathcal{S}] > 0$, we will use the notation $X_{\mathcal{S}}$ as the random variable distributed as $\mathbf{X}$ conditioned on $\mathbf{X} \in \mathcal{S}$.

Let $\mathcal{Y}$ be an interval in $\mathbb{R}$, we use $\mathcal{F}(\mathcal{Y})$ denote the set of all CDFs on $\mathcal{Y}$. We use $d : \mathcal{F}([0,1]) \times \mathcal{F}([0,1]) \to \mathbb{R}$ to denote a distance function between two CDFs on $[0,1]$.

$$d_{W1}(\mathbb{F}, \mathbb{F}') = \int_{r=0}^{1} |\mathbb{F}(r) - \mathbb{F}'(r)| dr$$

A **probability forecaster** is a function $h : \mathcal{X} \to \mathcal{F}(\mathcal{Y})$ that maps an input $x \in \mathcal{X}$ to a continuous CDF $h[x]$ over $\mathcal{Y}$. Note that $h[x]$ is a CDF, i.e. it is a function that takes in $y \in \mathcal{Y}$ and returns a real number $h[x](y) \in [0,1]$. We use $[\cdot]$ to denote function evaluation for x and $(\cdot)$ for y.

Let $\mathcal{H} \overset{\text{def}}{=} \{h : \mathcal{X} \to \mathcal{F}(\mathcal{Y})\}$ be the set of possible probability forecasters. We consider **randomized** forecasters $\mathbf{H}$ which is a random function taking values in $\mathcal{H}$.

Preliminary: Forecaster and Calibration

To clarify notation, $\mathbf{H}[\mathbf{X}](\mathbf{Y}), \mathbf{H}[x](\mathbf{Y})$ and $\mathbf{H}[x](y)$ are all random variables taking values in $[0,1]$, but they are random variables on different sample spaces.

- $\mathbf{H}[\mathbf{X}](\mathbf{Y})$ is a random variable on the sample space $\mathcal{H} \times \mathcal{X} \times \mathcal{Y}$ — All of $\mathbf{H}, \mathbf{X}, \mathbf{Y}$ are random.
- $\mathbf{H}[x](\mathbf{Y})$ is a random variable on the sample space $\mathcal{H} \times \mathcal{Y}$, while x is just a fixed value in $\mathcal{X}$.
- $\mathbf{H}[x](y)$ is a random variable on the sample space $\mathcal{H}$, while x, y are just fixed values in $\mathcal{X} \times \mathcal{Y}$.

Given some input $x \in \mathcal{X}$, an ideal forecaster should always output the CDF of the true conditional distribution $\mathbb{F}_{\mathbf{Y}|x}$. We call such a forecaster a "perfect forecaster".

Limits of Deterministic Forecaster

- Perfectly calibrated forecaster要求在输出空间上预测的CDF(概率分布)与真实的CDF一致；

- 但是，对于一个输入x，训练集中往往只有一个label y，仅通过一个label我们无法全面学习整个输出空间上的CDF；

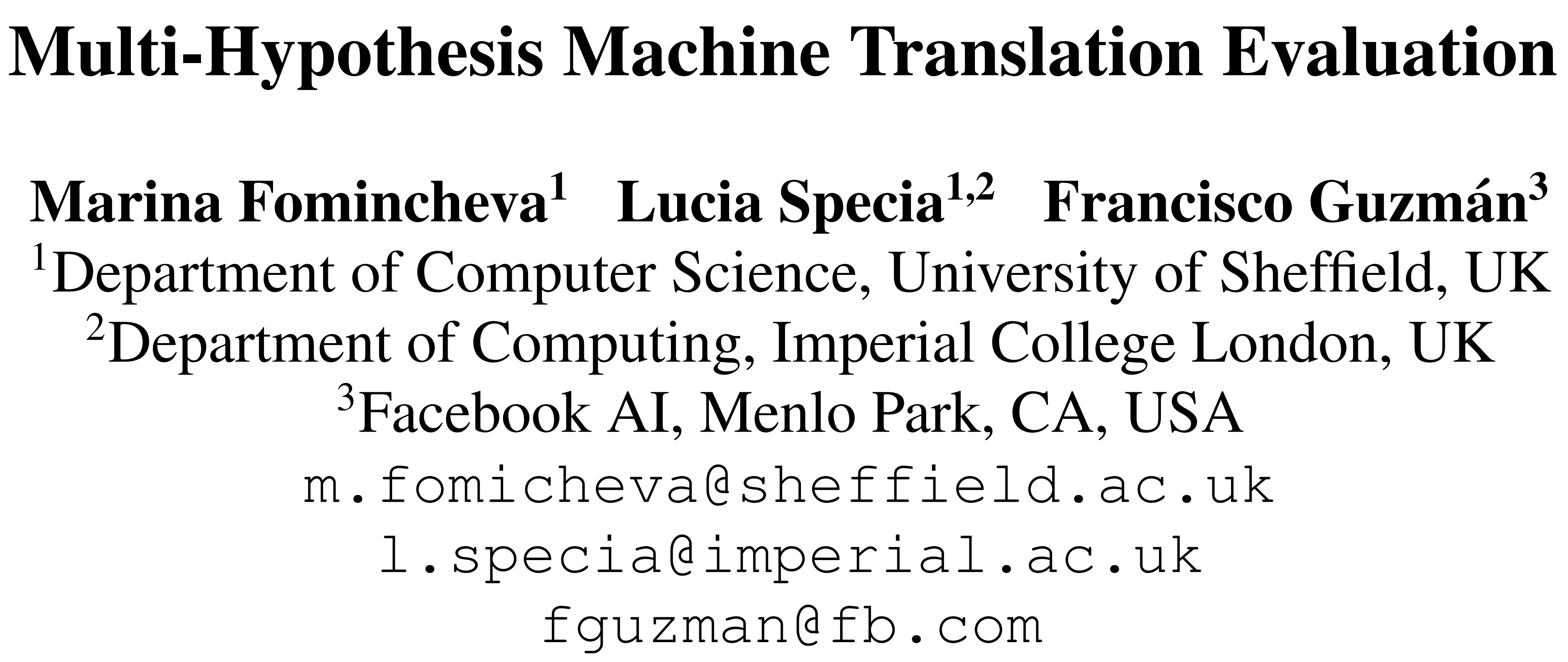
- 在NLP任务中miscalibration的来源也是如此，由于训练数据是single reference，我们的模型无法学习同义词、语序不同含义相同等重要语言学特征，从而导致模型在接近训练集的样例上over-estimate，在远离训练集的样例上under-estimate.

Limits of Deterministic Forecaster

- 训练集中，对于相同的x，我们只有一个y，但是研究calibration需要在一个样本集合上进行统计；

- 本文的思路“出其不意，但又意料之中”，既然只有一个label，那我就构建多个prediction，这样也能研究individual的样本的calibration.

- ACL2020也有一篇思路类似的文章



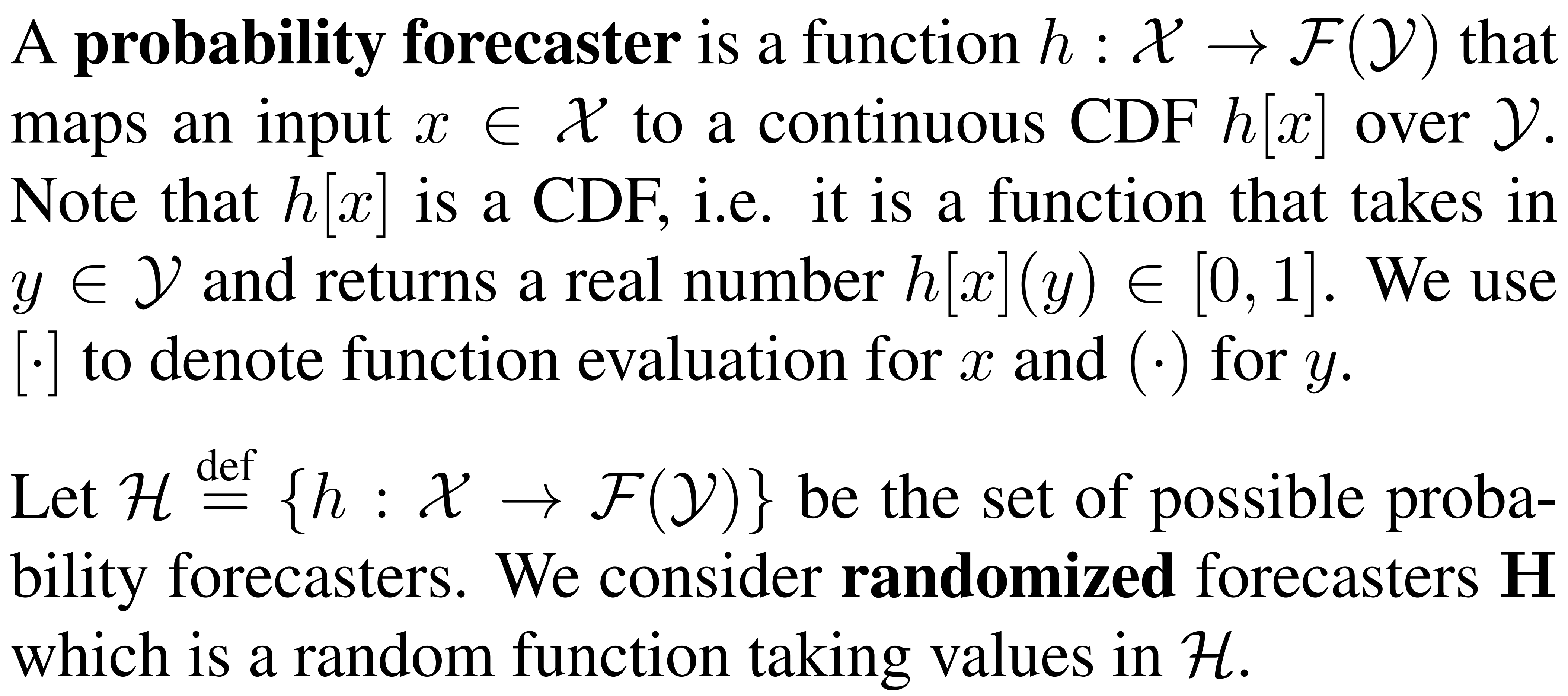
Multi-Hypothesis Machine Translation Evaluation

Marina Fomincheva[1] Lucia Specia[1,2] Francisco Guzmán[3]
[1]Department of Computer Science, University of Sheffield, UK
[2]Department of Computing, Imperial College London, UK
[3]Facebook AI, Menlo Park, CA, USA
m.fomicheva@sheffield.ac.uk
l.specia@imperial.ac.uk
fguzman@fb.com

Randomized Forecaster

- 因此我们可以理解作者为什么要randomize their forecaster.

- 如何训练randomized forecaster?



A **probability forecaster** is a function $h : \mathcal{X} \to \mathcal{F}(\mathcal{Y})$ that maps an input $x \in \mathcal{X}$ to a continuous CDF $h[x]$ over $\mathcal{Y}$. Note that $h[x]$ is a CDF, i.e. it is a function that takes in $y \in \mathcal{Y}$ and returns a real number $h[x](y) \in [0, 1]$. We use $[\cdot]$ to denote function evaluation for x and $(\cdot)$ for y.

Let $\mathcal{H} \overset{\text{def}}{=} \{h : \mathcal{X} \to \mathcal{F}(\mathcal{Y})\}$ be the set of possible probability forecasters. We consider **randomized** forecasters $\mathbf{H}$ which is a random function taking values in $\mathcal{H}$.

Summary

- Average calibration无法保证模型在特定的group甚至individual的样本上有准确的概率估计；

- 由于目前的训练数据对于输入x只有一个label y，对于deterministic forecaster无法进行individual calibration的研究；

- 为此我们需要使用randomized forecaster，对于确定的输入x我们多次输出是不同的。

Perfect Forecaster

- 对于给定的输入x，forecaster预测的概率分布$\mathbf{H}[x](\mathbf{Y})$与Y真实的条件分布$F_{\mathbf{Y}|x}$在输出空间上处处相等：

$$d_{W1}(\mathbb{F},\mathbb{F}') = \int_{r=0}^{1} |\mathbb{F}(r) - \mathbb{F}'(r)|dr$$

- 此时：$F_{\mathbf{H}[x](\mathbf{Y})}(c) = Pr(\mathbf{H}[x](\mathbf{Y}) \leq c) = Pr(F_{\mathbf{Y}|x}(\mathbf{Y}) \leq c) = c, \forall c \in [0,1]$

- 也即，我们的forecaster所预测的概率分布$(\mathbf{H}[x](\mathbf{Y}))$的分布$(F_{\mathbf{H}[x](\mathbf{Y})}(c))$是一个均匀分布

Perfect Forecaster

- 画图解释：$F_{\mathbf{H}[x](\mathbf{Y})}(c) = Pr(\mathbf{H}[x](\mathbf{Y}) \leq c) = Pr(F_{\mathbf{Y}|x}(\mathbf{Y}) \leq c) = c, \forall c \in [0,1]$

Randomized Forecaster

- 引入随机种子$\mathbf{R}$，$\mathbf{R}$在$[0, 1]$上的服从均匀分布；

- 定义$\mathbf{H}[x] = h[x, \mathbf{R}]$，通过对随机种子$\mathbf{R}$的采样来实现forecaster $\mathbf{H}[x]$的随机化；

- 将$h[x, \mathbf{R}](\mathbf{Y})$服从均匀分布的条件弱化为$h[x, \mathbf{R}](y)$服从均匀分布；注：本篇论文唯一一个没有严格证明的步骤，个人感觉这里不太严谨

- 作者采取了一种最简单的形式：$h[x, r](y) = r$

Training of Randomized Forecaster

- Calibration loss:

$$\mathcal{L}_{\text{PAIC}}(\theta) = \frac{1}{n}\sum_{i=1}^{n}|\bar{h}_\theta[x_i, r_i](y_i) - r_i|$$

- Concentration loss:

$$\mathcal{L}_{\text{NLL}}(\theta) = -\frac{1}{n}\sum_{i=1}^{n}\log\frac{d}{dy}\bar{h}[x_i, r_i](y_i)$$

- Total Loss:

$$\mathcal{L}_\alpha(\theta) = (1-\alpha)\mathcal{L}_{\text{PAIC}}(\theta) + \alpha\mathcal{L}_{\text{NLL}}(\theta)$$

Training of Randomized Forecaster

- 实验效果：

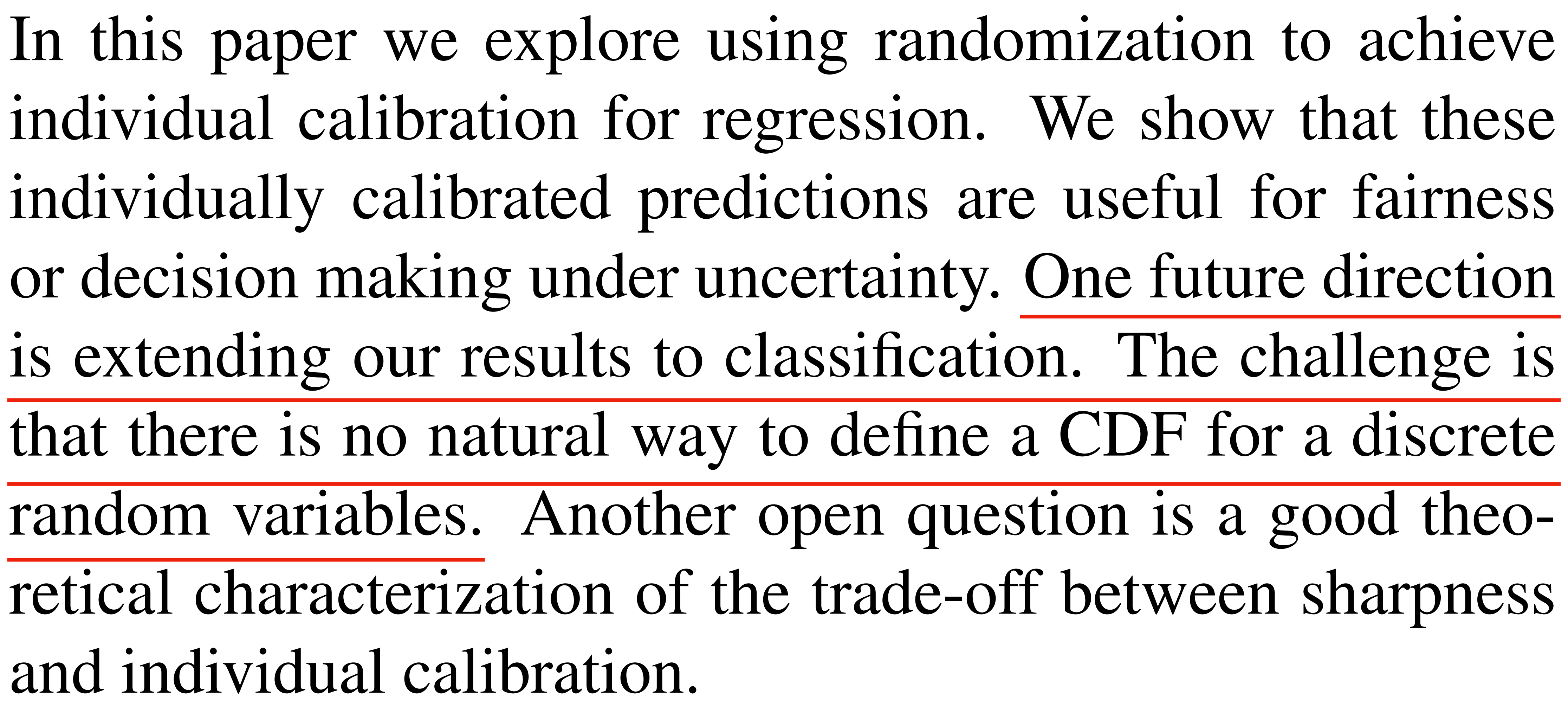
negative log likelihood
0.2
0.0
−0.2
−0.4
−0.6
−0.8
−1.0
0.1
0.2
0.3
0.4
calibration err - worst group
average stddev
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
calibration err - worst group
negative log likelihood
0.05
0.10
0.15
calibration err - interpretable group
average stddev
calibration err - interpretable group
α=0.06
α=0.08
α=0.10
α=0.15
α=0.20
α=0.40
α=0.70
α=1.00

- Calibration和NLL呈现一个trade-off的态势.

Training of Randomized Forecaster

- 在sequence generation的任务中，calibration和performance不一定是trade-off, 尽管calibration可能会影响单个token的accuracy，但是因为sequence generation过程存在rerank，calibrated probability有助于模型在search space中找到更好的candidate.

Conclusion and Future Work



In this paper we explore using randomization to achieve individual calibration for regression. We show that these individually calibrated predictions are useful for fairness or decision making under uncertainty. One future direction is extending our results to classification. The challenge is that there is no natural way to define a CDF for a discrete random variables. Another open question is a good theoretical characterization of the trade-off between sharpness and individual calibration.

Thanks