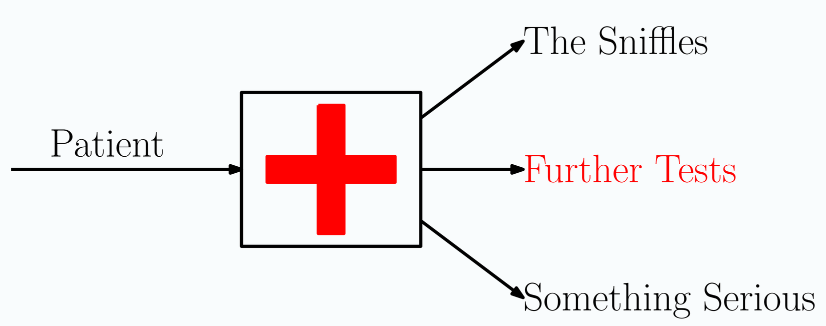
Selective Classification via One-Sided Prediction

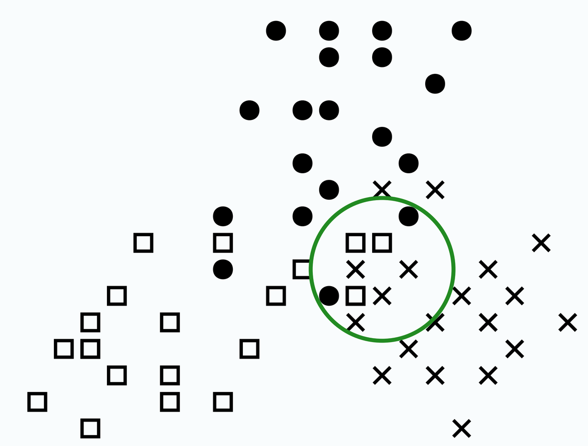
Aditya Gangrade,
Anil Kag,
Venkatesh Saligrama

Boston University

Selective Classification - what and why

Classification with the option to say 'I don't know', or '**reject**' a query.
- Dynamically collect features.
- Invoke Experts/human fallback





Goal: **Maximise coverage**, **while making total error smaller than a given level.**

Target error level $\varepsilon \ll 1$.
- Two key challenges
  - *Statistical* no supervision on what to reject.
  - *Computational* inherent non-convexity that makes training hard.

Prior Work

Naïve Scheme - Standard classifier f, and a post-hoc uncertainty score $\mathcal{U}$.
- Reject if $\mathcal{U}(f(x))$ is too big. Otherwise output $f(x)$.
- Using $\mathcal{U}(f(x)) = 1 - \max_k(f_k(x))$ is near-SOTA for DNNs [GEY17].

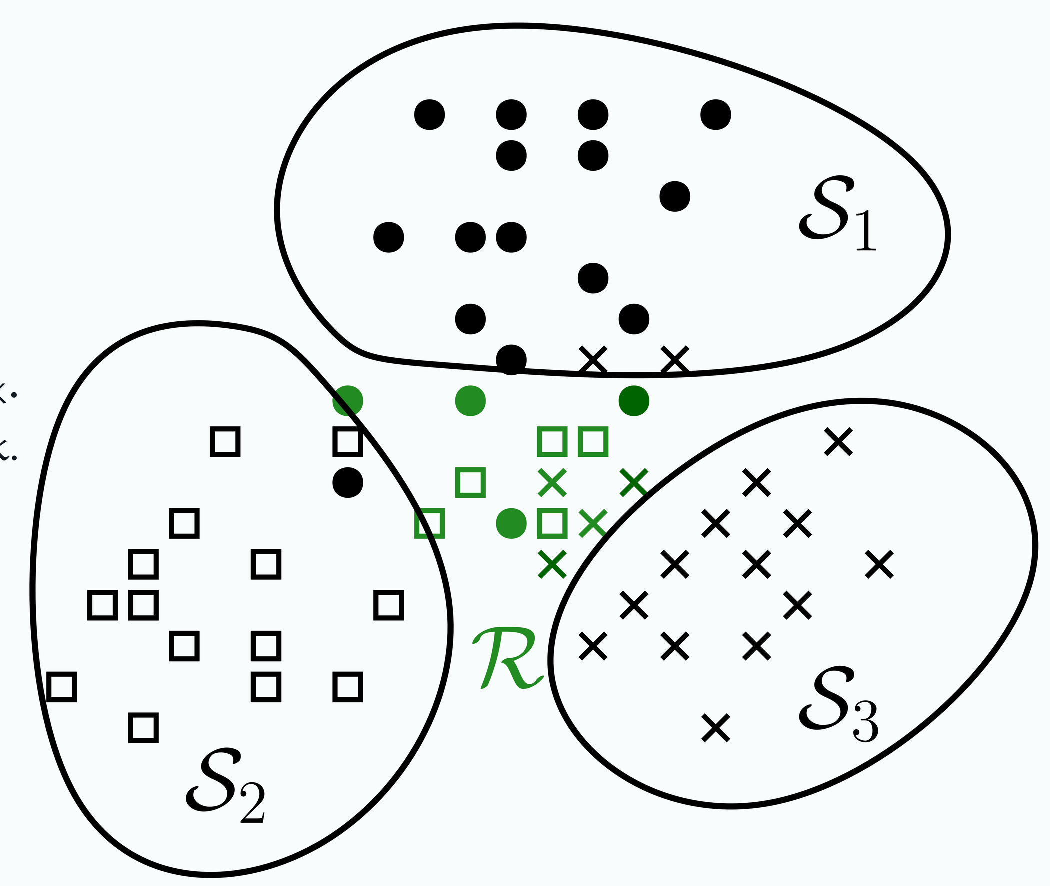
Gating Formulation - *Jointly train* a **gate** γ and a classifier f. [CDM16; EYW10; WEY11].
- If $\gamma(x) = 1$, reject. Else classify according to $f(x)$.
- Many approaches - relaxations & surrogates, alternating minimisation, architectures.
- SOTA: Selective Net [GEY19], Deep Gamblers [Liu+19].

(Lots of other interesting work - see paper)

Key Challenge - Gating SOTA $\approx$ Naïve SOTA.

Structure of Formulation

- Supervision $(X_i, Y_i) \overset{\text{i.i.d.}}{\sim} \mathbb{P}$; $Y_i \in [1:K]$.
- *Disjoint* decision sets $\mathcal{S}_1, \mathcal{S}_2, \dots, \mathcal{S}_K$.
- $\mathcal{S}_k$ = the decision region for class k.
- **Rejection region** $\mathcal{R} = \bigcap \mathcal{S}_k^c$.
- **Coverage**: $\sum \mathbb{P}(X \in \mathcal{S}_k)$.
- **Error**: $\sum \mathbb{P}(X \in \mathcal{S}_k, Y \neq k)$.



Formulation

- **Maximise Coverage**, keeping **error** smaller than ε.
- $\varepsilon \ll 1$ - 'Target error level'.
- Disjointness constraint to yield a valid classifier.

Enforcing disjointness makes training hard.

$$\max_{\{\mathcal{S}_k\}_{k\in[1:K]}} \sum_k \mathbb{P}(X \in \mathcal{S}_k)$$
$$\text{s.t.} \quad \sum_k \mathbb{P}(X \in \mathcal{S}_k, Y \neq k) \leq \varepsilon,$$
$$\forall k \neq k', \mathbb{P}(\mathcal{S}_k \cap \mathcal{S}_{k'}) = 0.$$

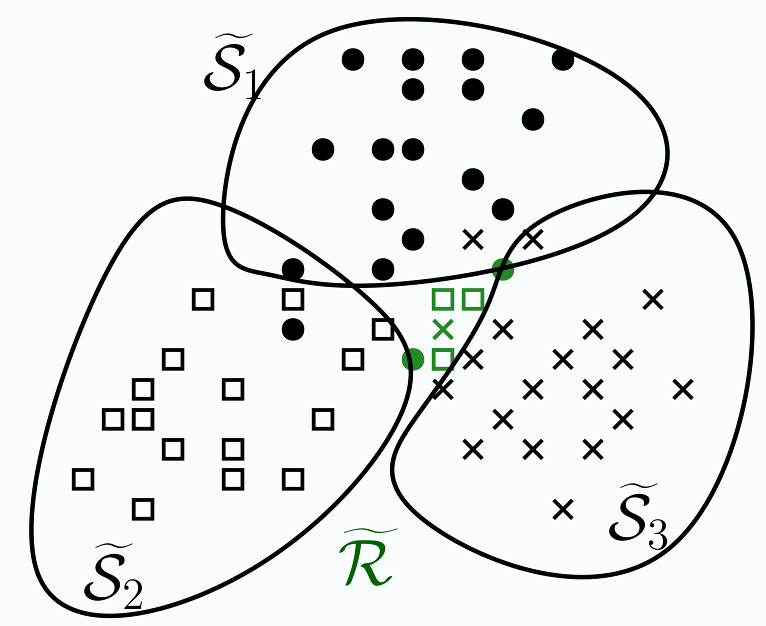
Relaxation to One-Sided Prediction

Drop the disjointness constraint.

$$\max_{\widetilde{\mathcal{S}}_k} \mathbb{P}(X \in \widetilde{\mathcal{S}}_k)$$
$$\text{s.t.} \quad \mathbb{P}(X \in \widetilde{\mathcal{S}}_k, Y \neq k) \leq \alpha_k \varepsilon.$$

1. The problem **decouples** into easier one-sided prediction (**OSP**) problems
   - Hyperparameters $\alpha_k \geq 0 : \sum \alpha_k = 1$.
2. Removing overlaps gives **feasible** $\{\mathcal{S}_k\}$ that are **near optimal**
   - $\sum \mathbb{P}(\mathcal{S}_k) \geq \mathrm{OPT}_\varepsilon - 2\varepsilon$.

1. $\implies$ Easy to train
2. $\implies$ not too lossy in the low ε regime.



Method Via Differentiable Relaxations

- Parametric class f^θ. $\quad f^\theta = (f_1^\theta, \dots, f_K^\theta)$.
- Solutions - threshold soft outputs. $\quad \widetilde{\mathcal{S}}_k^{\theta,t} = \{x : f_k^\theta(x) \geq t\}$.
- *Relaxed empirical* ***OSP***. $\quad \min_\theta L_k(\theta) \quad \text{s.t.} \quad C_k(\theta) \leq \varphi_k$
- **OSP** Lagrangian $\quad \mathcal{L}(\theta, \lambda_k) = L_k(\theta) + \lambda_k C_k(\theta)$.
- Train, select by matching λ_ks on validation.
  - Catch! Unviable due to hyperparameter search.

Solution:
- Heuristic: autotune using a minimax program.
- *Single parameter* μ to control total error.
- Joint Lagrangian

$$\mathcal{M}^\mu(\theta, \{\varphi_k\}, \{\lambda_k\}) = \sum_k (L_k(\theta) + \lambda_k(C_k(\theta) - \varphi_k) + \mu\varphi_k)$$

- Training:

$$\min_{\theta,\varphi} \max_{\lambda \geq 0} \mathcal{M}^\mu(\theta, \lambda, \varphi)$$

Empirical Performance

Comparison Against
- Naïve Softmax Response - f is a standard DNN classifier, reject if $\max_k f_k(x) < t$.
- Selective Net - DNN architecture for selective classification.
- Deep Gamblers - Loss function for selective classification based on gambling theory.

RESNET-32 models; CIFAR-10 dataset.

# of Samples Train.	Test.	Val.	Std. Error
45K	10K	5K	9.58%

- **Low Target Error Regime**

Target	OSP (Ours)		Naïve		SelectiveNet		Deep Gamblers	
Error	Cov.	Error	Cov.	Error	Cov.	Error	Cov.	Error
2%	**80.6**	1.91	75.1	2.09	73.0	2.31	74.2	1.98
1%	**74.0**	1.02	67.2	1.09	64.5	1.02	66.4	1.01
0.5%	**64.1**	0.51	59.3	0.53	57.6	0.48	57.8	0.51

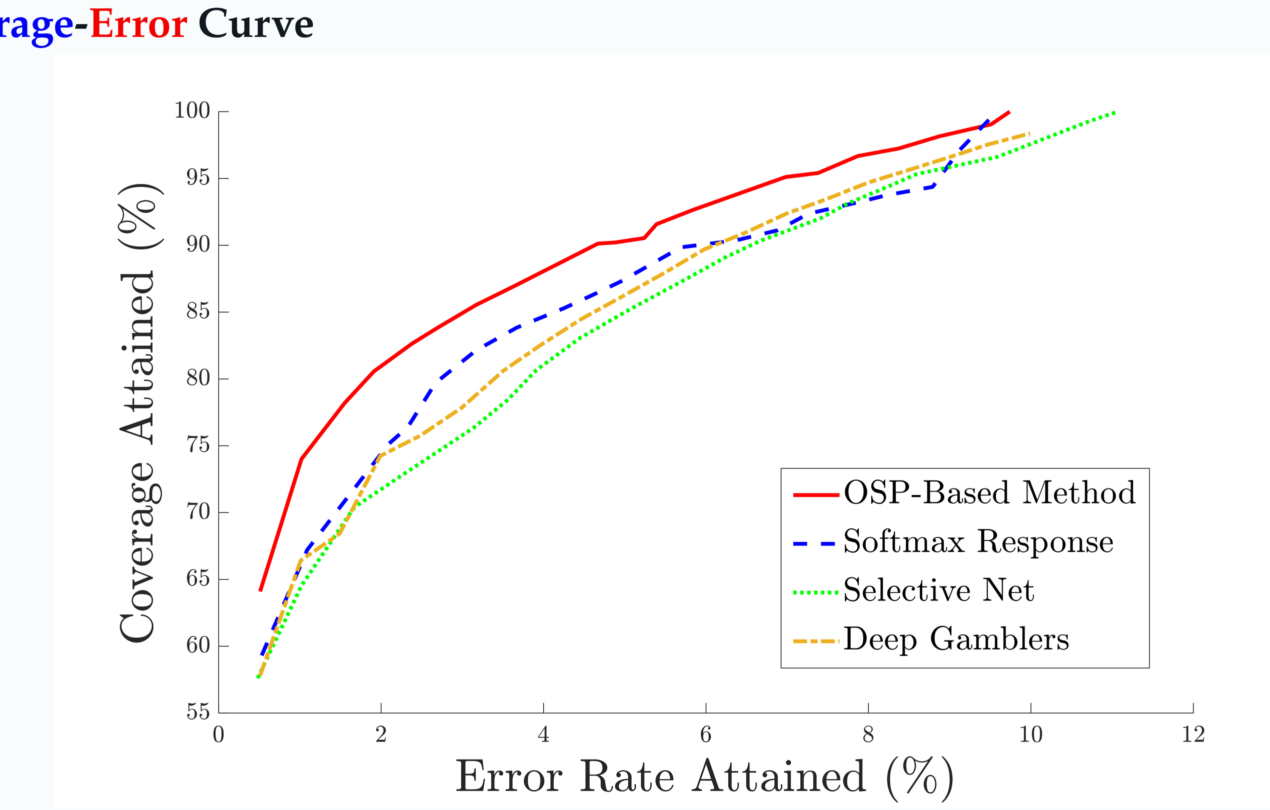
- Gains of at least **4.5**% against best competitor.
- **OSP-based** $\gg$ Naïve SOTA

- **High Target Coverage Regime**
  - 'Dual' formulation - minimise error subject to large coverage.

Target	OSP (Ours)		Naïve		SelectiveNet		Deep Gamblers	
Cov.	Cov.	Error	Cov.	Error	Cov.	Error	Cov.	Error
100%	100	9.74	100	**9.58**	100	11.07	100	10.81
95%	95.1	**6.98**	95.2	8.74	94.7	8.34	95.1	8.21
90%	90.0	**4.67**	90.5	6.52	89.6	6.45	90.1	6.14

- Outside of very high coverage, OSP-based still outperforms by > 1%.
- Very surprising, given design.

- **Coverage-Error Curve**



- **Overlap Characteristics**
  - Empirical Total Overlap is much smaller than 2ε.

Target Error (%)	2	1	0.5
Empirical Total Overlap (%)	0.09	0.01	0.00

LaTeX TikZposter