Robust Satisficing Ensemble: Scalable Model Aggregation Under Distribution Shifts

Ahmet Faruk Cetinkaya* Enes Agirman* Cem Tekin

Bilkent University

a.cetinkaya3@gatech.edu, enes.agirman@bilkent.edu.tr, cemtekin@ee.bilkent.edu.tr

NeurIPS 2026

Outline

Motivation

Robust Satisficing

The RSE Algorithm

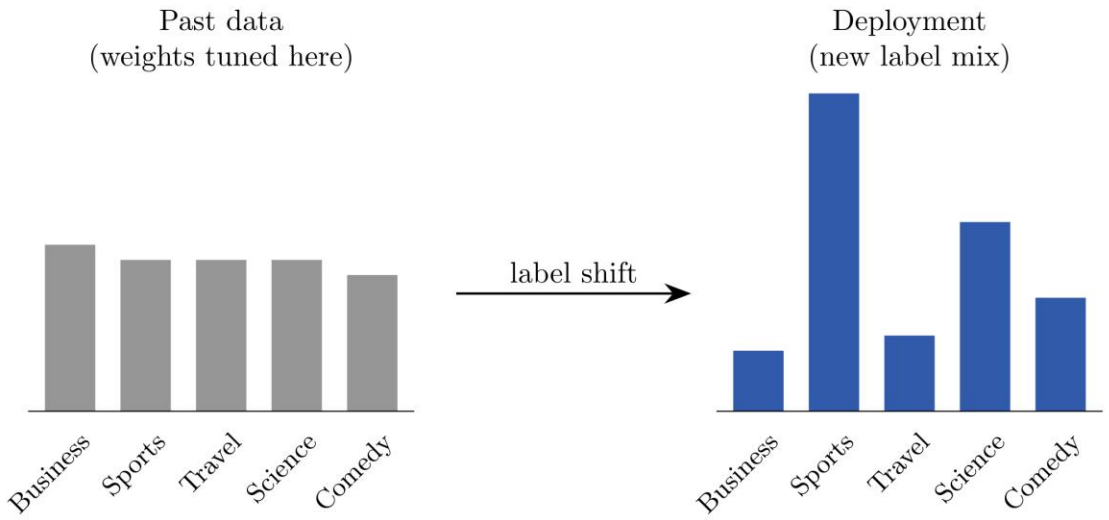
Experimental Results

Conclusion

Motivation

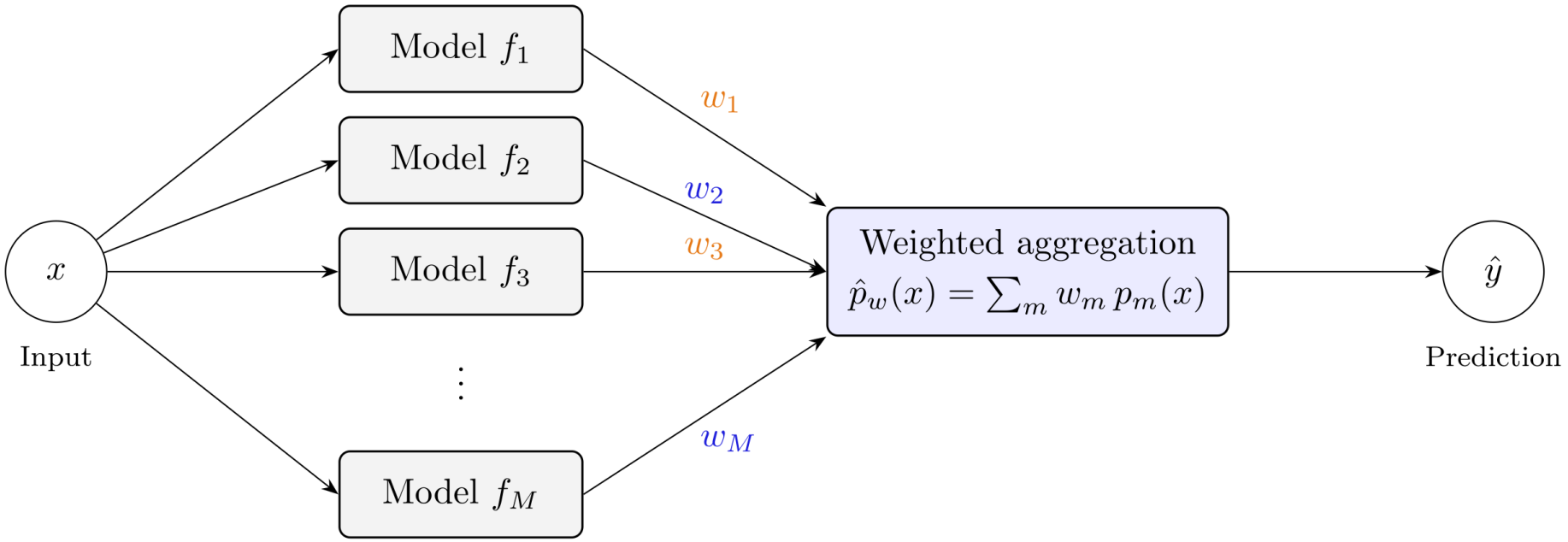
- Ensembles combine diverse pre-trained models
- Combination tuned on a single calibration distribution
- Under distribution shift, models often fail together
- Need: aggregation that is reliable under shift and scalable

Motivating Example



- LLM ensemble labels news headlines
- Tuned on past data; topic mix keeps changing
- Accuracy must stay above an acceptable level

Problem Setting



- Pre-trained models stay fixed; only the weights are learned
- Goal: weights reliable under distribution shift

Outline

Motivation

Robust Satisficing

The RSE Algorithm

Experimental Results

Conclusion

Robust Satisficing

Fragility

Fix a target risk τ; minimize the fragility

$$\kappa_\tau(w) = \min\left\{\kappa \geq 0 : \;\; R(w,q) \leq \tau + \kappa\, D_{\mathrm{KL}}(q \,\|\, q_{\mathrm{ref}}) \quad \forall q \in \Delta_K\right\}$$

- Small κ: risk grows slowly under shift
- No need to know the shift size in advance

Outline

Motivation

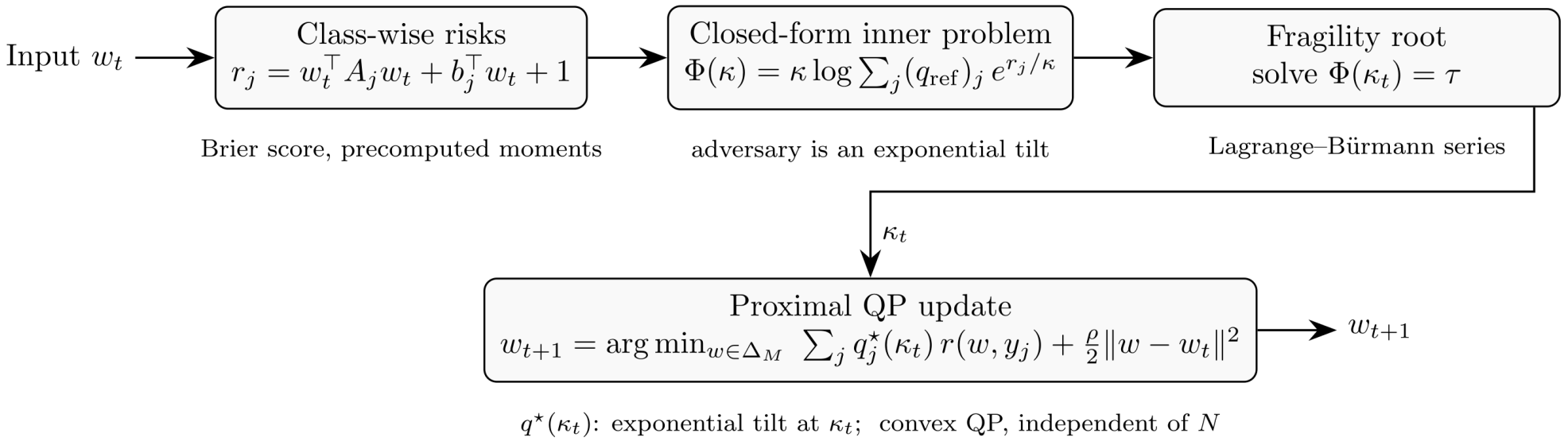
Robust Satisficing

The RSE Algorithm

Experimental Results

Conclusion

The RSE Algorithm



Goal: $\min_{w \in \Delta_M} \kappa_\tau(w)$

- KL + RS: exponential-tilt adversary
- Lagrange–Bürmann: fully analytical inner solution
- Each QP step provably decreases fragility, independent of dataset size

Outline

Motivation

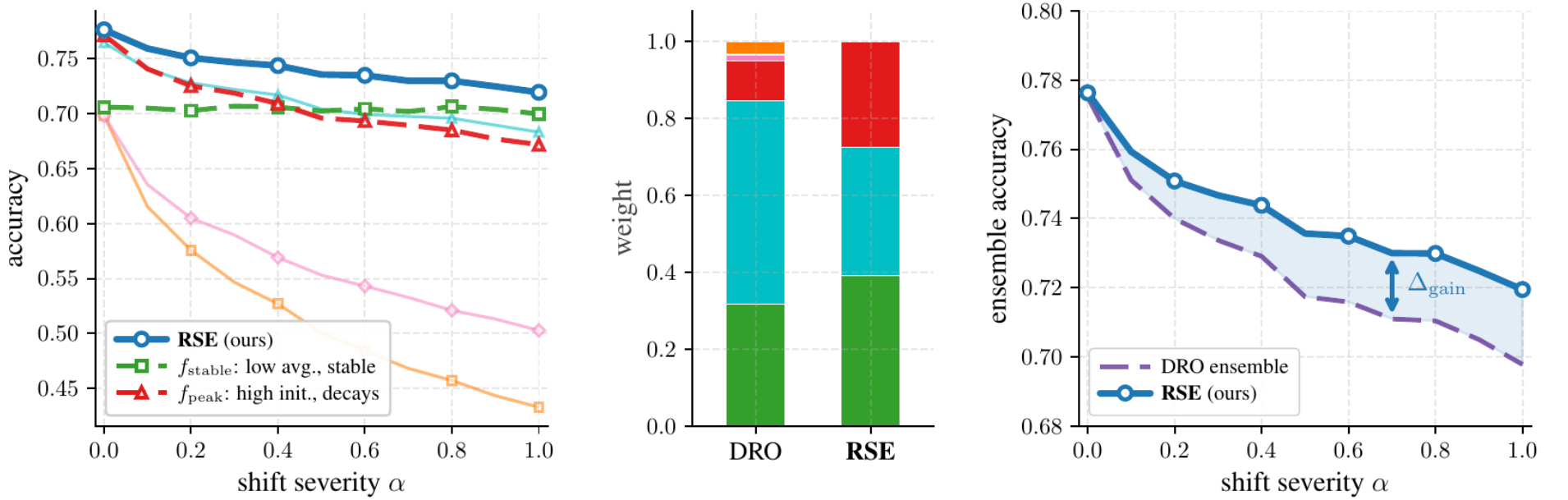
Robust Satisficing

The RSE Algorithm

Experimental Results

Conclusion

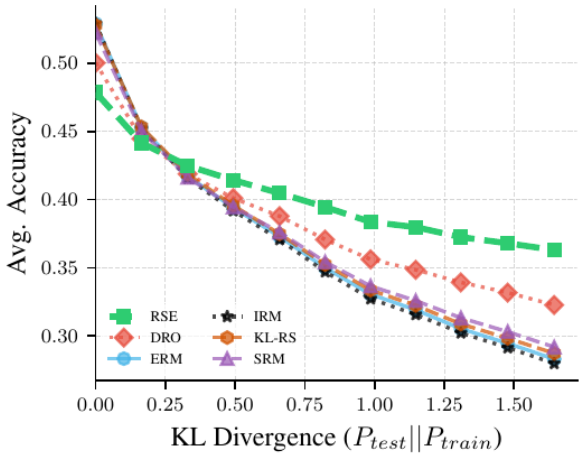
Why RSE Works: Comparison with DRO



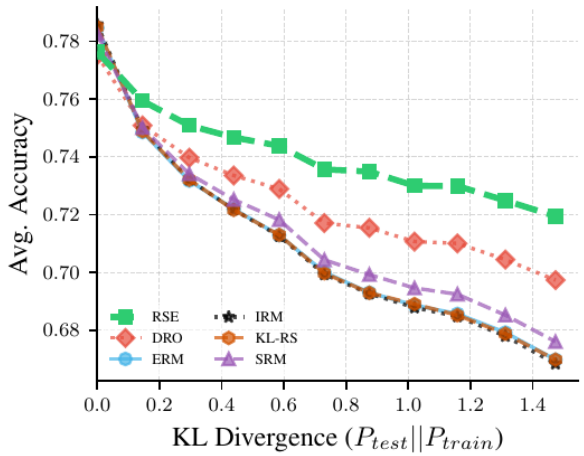
(a) Individual model profiles (b) Weight assignment (c) Ensemble performance under shift

- DRO: worst case only inside a fixed ball of radius ρ
- DRO leans on one model; RSE balances stable and high-peak models

Experiments: Adversarial Label Shift



(a) SST-5 Accuracy



(b) TREC Accuracy

Most severe shift (accuracy %)

Method	**SST-5**	**TREC**
Uniform	25.1	61.7
ERM	28.3	67.1
KL-RS	28.8	66.9
SRM	29.3	67.7
IRM	28.0	66.7
DRO	32.5	69.9
RSE (ours)	**36.3**	**72.0**

- 8 open-weight LLMs; labels shifted toward the hardest class
- RSE: slowest degradation, **best under severe shift**

Experiments: Natural Distribution Shifts

- **CivilComments:** shift across demographic groups
- **HuffPost:** shift over time
- RSE: best on every metric

	CivilComments		HuffPost	
Method	**Acc (%)**	**Worst group**	**Acc (%)**	**Worst year**
ERM	70.1	56.8	78.1	76.8
DRO	70.6	57.7	76.3	75.6
IRM	69.8	56.5	78.4	77.0
SRM	68.8	55.1	78.4	77.0
KL-RS	69.6	56.3	78.1	76.7
RSE (ours)	**72.1**	**59.7**	**78.8**	**77.2**

Conclusion

- RSE: ensemble weights that minimize fragility
- Fast, scalable, with theoretical guarantees
- Outperforms robust baselines under label, group and temporal shifts

Thank You!

github.com/Bilkent-CYBORG/RSE

This work was supported by TUBITAK under Grant No. 124E065, TUBITAK 2024 Incentive Award, and TUBA-GEBIP 2023 Award.