Distributionally Robust Language Modeling
Yonatan Oren*, Shiori Sagawa*, Tatsunori B. Hashimoto*, Percy Liang

**Goal**
Given multi-domain training text, learn a language model that performs well on unknown test distributions.

**Problem**
With standard training, a model performs worse with more data from outside the target domain.

**Tool: Distributionally Robust Optimization**
To achieve low loss on an unknown test distribution, we optimize the loss on the worst-case test distribution.

\[ \mathbb{E}_{\mathcal{P}}[\ell(x; \theta)] \leq \sup_{\mathcal{P}_x \in \mathcal{P}} \mathbb{E}_{\mathcal{P}_x}[\ell(x; \theta)] \]

if \( P_x^{\text{test}} = \) is in \( \mathcal{P} = \{ \ldots \} \)

unknown test distribution

uncertainty set: set of potential distributions

**Idea 1: Topic-Based Uncertainty Sets**
Problem: Naïve, sentence-based uncertainty sets are too conservative. No domain information.

Solution: Define the uncertainty set by latent topics.

\[ \mathcal{P} = \{ \ldots \} \]

uncertainty set: mixtures of latent topics

Result: Topics improve Yelp perplexity

**Idea 2: Baselined Loss**
Problem: DRO on NLL loss overemphasizes hard topics

Solution: Define a loss baselined by topic difficulty

\[ \ell((x, z); \theta) = -\log p_{\theta}(x) + H(z) \]

Accounts for topic difficulty

Optimizes the distribution fit between the worst-case topic \( z \) and model,

\[ \text{KL}(p_{x|z} \parallel p_{\theta}) \]

We estimate topic difficulty:

\[ H(\tilde{H}) \]

resulting in lower perplexity.

**Experiments**
Set-up: Train a Transformer on mixture of Yelp \( (\alpha^*) \) and One Billion Words \( (1 - \alpha^*) \).

Minority training domain (Yelp): Topic DRO improves Yelp perplexity, estimating the optimal Yelp vs. news trade-off. Pure Yelp results (32) are impossible for our setting due to unknown test domain.

**References**
DRO: Ben-Tal+ 2013, Rockafellar and Uryasev 2000, Duchi and Namkoong 2018
Topics in DRO: Hu+ 2018