Distributionally Robust Language Modeling

Yonatan Oren*[1]   Shiori Sagawa*[1]   Tatsunori B. Hashimoto*[1,2]   Percy Liang[1]
(* equal contribution)
[1]Stanford Computer Science   [2]Stanford Statistics
{yonatano,thashim}@stanford.edu   {ssagawa,pliang}@cs.stanford.edu

Abstract

Language models are generally trained on data spanning a wide range of topics (e.g., news, reviews, fiction), but they might be applied to an a priori unknown target distribution (e.g., restaurant reviews). In this paper, we first show that training on text outside the test distribution can degrade test performance when using standard maximum likelihood (MLE) training. To remedy this without the knowledge of the test distribution, we propose an approach which trains a model that performs well over a wide range of potential test distributions. In particular, we derive a new distributionally robust optimization (DRO) procedure which minimizes the loss of the model over the *worst-case* mixture of topics with sufficient overlap with the training distribution. Our approach, called *topic conditional value at risk (topic CVaR)*, obtains a 5.5 point perplexity reduction over MLE when the language models are trained on a mixture of Yelp reviews and news and tested only on reviews.

1   Introduction

Large-scale language modeling plays a central role in both text generation (Sordoni et al., 2015; Nallapati et al., 2016) and unsupervised pre-training (Vaswani et al., 2013; Dai and Le, 2015; McCann et al., 2017; Peters et al., 2018; Devlin et al., 2018; Radford et al., 2018). In both settings, a single language model is trained on a large corpus containing a range of topics (e.g. news, fiction, and reviews). This language model is then applied in many different tasks, each with a specific test distribution (e.g., analyzing the sentiment of restaurant reviews). Can we train a single general-purpose language model that works across a wide range of potential test distributions?

In this work, we first demonstrate that standard maximum likelihood training on a large, heterogeneous dataset can fail to achieve this goal.
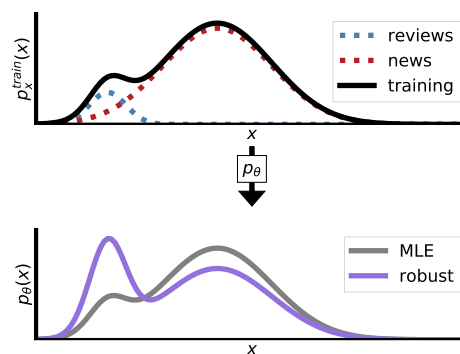

Figure 1. Illustration of a training corpus as a density (black) with mostly news stories (red) and a small number of restaurant reviews (blue). The standard MLE model (gray) reflects the underlying data and assigns little weight to reviews, and thus performs poorly on reviews. A more robust model should try to equalize the weight across all topics so that it can perform well regardless of which topics appear at test time.

While more data is generally better, the presence of text outside the target distribution actually *degrades* performance on a target test distribution. For example, a language model trained on Yelp reviews achieves a perplexity of 32, and this perplexity increases to 43 when trained on a mixture of 10% Yelp and 90% newswire sentences from the One Billion Word Benchmark (Chelba et al., 2013). Performance degrades because existing maximum likelihood estimation (MLE) objectives tend to emphasize model performance on more common sentences and topics at the expense of infrequent ones (Figure 1).

While the above performance degradation can be mitigated by fine-tuning and domain adaptation techniques (Shimodaira, 2000; Quiñonero-Candela et al., 2009; Daume III, 2007; Ben-David et al., 2010; Blitzer et al., 2011; Pryzant et al., 2017; Ganin and Lempitsky, 2015; Tzeng et al., 2014), these methods require knowing the test distribution and training a separate model specific to each target distribution. Instead, we aim to train a

single model that performs well across many unknown test distributions.

In order to do this, we will train a model that performs uniformly well over an entire family of potential test distributions. Since we cannot expect to do well on all possible test distributions, we consider the *subpopulation shift* setting, in which the test distribution is a subpopulation of the training distribution, and seek good performance across all such test distributions (e.g. Yelp reviews in a Yelp-newswire mixture). [1] In other words, adding data from topics outside the test topics should not hurt. It seems reasonable to protect against subpopulation shifts, intuitively because large-scale data collection schemes are designed to cover a diverse array of topics as a way to generalize to potential test distributions.

We train a model that performs well over all subpopulations by minimizing the risk for the *worst-case* subpopulation, following the distributionally robust optimization (DRO) literature (Ben-Tal et al., 2013). While an existing DRO framework called the conditional value at risk (CVaR) ensures uniformly good performance across subpopulations (Rockafellar and Uryasev, 2000; Duchi and Namkoong, 2018), we demonstrate that naïvely applying this approach to language modeling fails due to three challenges. First, the existing CVaR approach is too conservative because it considers robustness to *arbitrary* subpopulations. Such worst-case subpopulations are attained by adversarially choosing the hardest, most unusual sentences. Instead, we propose to consider *meaningful* subpopulations, defined by topics in a corpus (Hu et al., 2018). Second, applying CVaR directly to log loss results in a loss which is biased towards topics with high entropy, instead of those for which the model performs poorly relative to what is possible. We correct this by introducing a new *baselined* loss function which measures losses relative to the entropy of each topic. Finally, existing optimization algorithms for CVaR are either inapplicable to topic-based robustness sets or unscalable because they require batch optimization. We develop a scalable online algorithm which identifies the worst-performing topics at each iteration and upweights examples from those topics.

With these methodological improvements, we

---

[1] The subpopulation assumption refers to overlaps in *distributions*, rather than individual examples. Our assumptions do not require overlap in the training and test data.

demonstrate that our approach, *topic CVaR*, improves robustness against subpopulation shifts. Topic CVaR reduces perplexity on the Yelp review corpus by 5.5 points compared to MLE when trained on the Yelp-One Billion Word Benchmark mixture from before. We also show improved robustness even when the shift is not strictly a subpopulation shift. Topic CVaR also achieves a 4 point perplexity reduction on a test distribution (TripAdvisor hotel reviews) that is similar to, but not strictly a subpopulation of the training distribution (Yelp and newswire text).

2 Problem Statement

Our goal is to learn a language model p_θ based on sentences sampled from the training distribution $x \sim p_{\mathrm{x}}^{\mathrm{train}}$, such that p_θ performs well on unknown test distributions $p_{\mathrm{x}}^{\mathrm{test}}$.

Language models p_θ are generally trained to approximate $p_{\mathrm{x}}^{\mathrm{train}}$ by minimizing the KL divergence $\mathrm{KL}\left(p_{\mathrm{x}}^{\mathrm{train}} \,\|\, p_\theta\right)$ via maximum likelihood estimation (MLE),

$$\inf_\theta \mathbb{E}\left[-\log p_\theta(x)\right]. \tag{1}$$

When $p_{\mathrm{x}}^{\mathrm{test}} = p_{\mathrm{x}}^{\mathrm{train}}$, classical statistical theory guarantees that a model trained via MLE performs well on the test distribution given sufficient data. However, when $p_{\mathrm{x}}^{\mathrm{test}}$ is not identical to $p_{\mathrm{x}}^{\mathrm{train}}$, MLE can perform poorly no matter how much data is observed. This is because the test set might consist solely of sentences from topics that are infrequent during training, to which MLE would assign low probabilities.

To illustrate this point, consider the toy example drawn in Figure 2. In this example, the training distribution $p_{\mathrm{x}}^{\mathrm{train}}$ is a multinomial distribution over six possible sentences A–F, with two from reviews and four from news. Sentence F is ungrammatical and thus has an extremely low probability. The training distribution includes 10% reviews and 90% news, whereas the test distribution could be all reviews, all news, or a mixture. MLE assigns low probabilities to any review and thus performs poorly when evaluated solely on reviews. To be robust, we intuitively need a more conservative objective that encourages models to assign higher probabilities to rare but valid sentences.

In order to achieve this, we want to learn a model p_θ which performs well in situations where $p_{\mathrm{x}}^{\mathrm{train}} \neq p_{\mathrm{x}}^{\mathrm{test}}$ for a large set of potential test distributions $\mathcal{P}$, termed the uncertainty set. By training
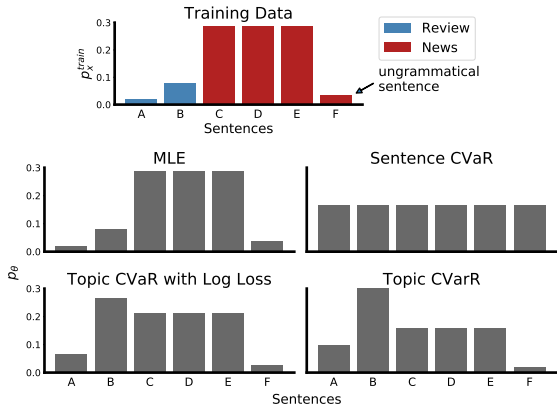
Figure 2. Toy example of a multinomial distribution over six sentences (top). Different panels illustrate models learned by different training procedures. MLE fits common topics (news) at the expense of rare ones (reviews). Sentence CVaR is too conservative, overemphasizing the ungrammatical sentence. Topic CVaR with log loss overemphasizes difficult topics (news) over easy ones (review). Topic CVaR (with baselining) balances the weight assigned to each topics, as desired.

a model that performs well on all distributions in the uncertainty set $\mathcal{P}$, we can ensure good test performance as long as $p_{\mathrm{x}}^{\mathrm{test}} \in \mathcal{P}$.

More formally, this approach falls under the framework of distributionally robust optimization (DRO) (Ben-Tal et al., 2013). With DRO, we optimize a model for loss ℓ and a set of potential test distributions $\mathcal{P}$ by minimizing the risk under the *worst-case* distribution in $\mathcal{P}$,

$$\sup_{p_{\mathrm{x}} \in \mathcal{P}} \mathbb{E}_{p_{\mathrm{x}}}[\ell(x; \theta)]. \quad (2)$$

Observe that the above worst-case objective does not depend on the unknown quantity $p_{\mathrm{x}}^{\mathrm{test}}$. The objective also upper bounds the test risk for all $p_{\mathrm{x}}^{\mathrm{test}} \in \mathcal{P}$ as

$$\mathbb{E}_{p_{\mathrm{x}}^{\mathrm{test}}}[\ell(x; \theta)] \leq \sup_{p_{\mathrm{x}} \in \mathcal{P}} \mathbb{E}_{p_{\mathrm{x}}}[\ell(x; \theta)], \quad (3)$$

so optimizing the above objective gives guarantees on test performance whenever $p_{\mathrm{x}}^{\mathrm{test}} \in \mathcal{P}$.

DRO provides a conceptually appealing framework for learning under train-test mismatch. However, it crucially depends on both the choice of uncertainty set $\mathcal{P}$ and loss ℓ, and we will discuss these choices in the next section.

3 Robust Language Modeling

We will begin by applying standard distributionally robust optimization approaches to the log loss (Section 3.1), and showing that this naïve approach suffers from two drawbacks:

1. Existing DRO uncertainty sets $\mathcal{P}$ are too conservative.

2. The log loss overemphasizes topics with inherently high entropy.

These drawbacks will motivate our development of a new approach we call topic CVaR, which addresses these two problems (Sections 3.2 and 3.3).

3.1 Robustness to arbitrary subpopulations

Observing that MLE is not robust because it assigns low probabilities (i.e. incurs high losses) on rare sentences, we might initially try to define $\mathcal{P}$ as individual training examples to ensure low loss on *all* data points. However, this is far too conservative, since the worst-case distribution would consist of exactly one data point. Therefore, we may want to optimize a slightly more realistic uncertainty set consisting of all sufficiently large subpopulations of the training distribution.

Minimizing losses over all subpopulations of the training distribution can be formulated as a type of distributionally robust optimization (DRO) problem (Duchi and Namkoong, 2018), which has been used to regularize models (Duchi and Namkoong, 2016), defend against adversarial examples (Sinha et al., 2018), and improve the fairness of models (Hashimoto et al., 2018).

One type of distributionally robust loss is known as conditional value at risk (CVaR) which guarantees low losses on all α-fraction subpopulations of the training distribution (Rockafellar and Uryasev, 2000). This corresponds to defining the uncertainty set $\mathcal{P}$ as all sentence distributions that are α-*covered* by $p_{\mathrm{x}}^{\mathrm{train}}$,

$$\mathcal{P}_{\mathrm{x}}^{\alpha} := \{p_{\mathrm{x}} : \alpha p_{\mathrm{x}}(x) \leq p_{\mathrm{x}}^{\mathrm{train}}(x) \ \forall x\}. \quad (4)$$

This is equivalent to defining $\mathcal{P}_{\mathrm{x}}^{\alpha}$ as the set of p_{x} which fulfills $p_{\mathrm{x}}^{\mathrm{train}} = \alpha p_{\mathrm{x}} + (1-\alpha)p_{\mathrm{x}}^{\mathrm{other}}$ for some distribution $p_{\mathrm{x}}^{\mathrm{other}}$.

To achieve low loss on all possible test distributions in $\mathcal{P}_{\mathrm{x}}^{\alpha}$, we minimize the expected loss under the worst-case distribution,

$$\sup_{p_{\mathrm{x}} \in \mathcal{P}_{\mathrm{x}}^{\alpha}} \mathbb{E}_{x \sim p_{\mathrm{x}}}[\ell(x; \theta)]. \quad (5)$$

For the remainder of the paper, we will refer to this approach as *sentence CVaR*, highlighting the fact that it considers robustness over arbitrary sets of sentences. It intuitively encourages uniform performance across all subpopulations of sentences

by downweighting sentences with low loss, and upweighting sentences with high loss.

Because sentence CVaR considers *arbitrary* groups of examples, it can be too conservative in our problem setting. While sentence CVaR can prevent modeling common sentences at the cost of rare ones, it can also encourage modeling invalid sentences at the expense of valid ones. Returning to our example in Figure 2 with $\ell(x;\theta) = -\log p_\theta(x)$, sentence CVaR with for sufficiently low α achieves perfectly uniform performance. It equalizes likelihoods across all sentences, which unfortunately also results in high probabilities assigned to the ungrammatical sentence F.

3.2 Robustness over Topics

Sentence CVaR is too conservative since it allows for arbitrary groups — including ones consisting of purely invalid sentences. To remedy this, we will optimize models for all *meaningful* subpopulations instead of *arbitrary* ones.

One way to achieve this is through robustness over topics, rather than individual examples. For example, a news corpus often contains a variety of topics (politics, business, opinion, food) and a test corpus may contain these topics with different proportions. A robust language model should perform well on a wide range of topic mixtures without taking the topic identity as an input.

Formally, we posit that each sentence x belongs to some latent topic z, which has a sentence distribution $p_{x|z}$. We want our models to be robust to shifts in the topic distribution, where we have $z \sim p_z^{\text{train}}$ and $z \sim p_z^{\text{test}}$. In this case, we can define a natural uncertainty set for CVaR, defined over latent topics rather than individual examples. Extending the definition of α-covered distributions to topics, we have the set

$$\mathcal{P}_z^\alpha := \{p_z : \alpha p_z(z) \le p_z^{\text{train}}(z) \ \forall z\} \quad (6)$$

and the objective is the expected loss under the worst-case topic distribution,

$$\sup_{p_z \in \mathcal{P}_z^\alpha} \mathbb{E}_{z \sim p_z}\left[\mathbb{E}_{x \sim p_{x|z}}\left[\ell(x;\theta)\right]\right]. \quad (7)$$

This objective intuitively encourages uniform loss across topics by upweighting topics incurring high losses and downweighting topics with low losses, while keeping the conditional distribution of sentences given a topic constant.

3.3 Baselined Loss Function

Recall that DRO depends critically on the choice of uncertainty set and loss function. Having specified the uncertainty set, we now turn to the choice of loss $\ell(x;\theta)$. While the log loss $\ell(x;\theta) = -\log p_\theta(x)$ is the standard choice in language modeling, we show that this approach has a flaw in the robust setting and propose a corrected loss.

Log Loss. Using log loss on CVaR encourages uniform *absolute* log-likelihoods across topics even if some topics are much harder than others. For example, consider a model which performs nearly optimally on difficult topics and highly suboptimally on easy topics. Since log loss measures absolute performance, it would force the model to focus on the difficult topic *even if the model can't improve further on this topic*. In the example in Figure 2, news is emphasized over reviews because news has higher entropy and thus higher difficulty. Empirically, we observe that log loss with CVaR forces the models to focus almost entirely on the difficult topics such as long news stories.

Baselined Loss. We now propose a new baselined loss, which encourages uniform *relative* performance across topics. We refer to our approach with the baselined loss as *topic CVaR*.

The baselined loss function $\ell(x,z;\theta) = \log p_{x|z}(x \mid z) - \log p_\theta(x)$ evaluates the performance of the model relative to the best possible model for the topic, $\log p_{x|z}(x \mid z)$. Although we do not observe $\log p_{x|z}(x \mid z)$, we will show later in section 4.2 that we can estimate sufficient statistics of $\log p_{x|z}(x \mid z)$ that allow us to compute the baselined loss. By using baselined loss, we intuitively encourage models to perform as well as it can on each topic while making optimal trade-offs among topics.

Plugging the baselined loss into the robust objective (7), the optimization problem is

$$\sup_{p_z \in \mathcal{P}_z^\alpha} \mathbb{E}_{z \sim p_z}\left[\mathbb{E}_{x \sim p_{x|z}}\left[\log p_{x|z}(x \mid z) - \log p_\theta(x)\right]\right],$$

$$(8)$$

which can be simplified to

$$\sup_{p_z \in \mathcal{P}_z^\alpha} \mathbb{E}_{z \sim p_z}\left[\text{KL}\left(p_{x|z} \,\|\, p_\theta\right)\right]. \quad (9)$$

Topic CVaR thus minimizes the per-topic KL divergences, and this interpretation fits nicely with

a general goal of training p_θ that matches the test distribution. Unlike in the MLE case, minimizing the KL is not equivalent to minimizing the log loss. In MLE, minimizing $\text{KL}(p_{\text{x}}^{\text{train}} \| p_\theta) = \mathbb{E}\left[\log p_{\text{x}}^{\text{train}}(x) - \log p_\theta(x)\right]$ is equivalent to minimizing the log loss because $\log p_{\text{x}}^{\text{train}}(x)$ can be treated as a constant. However, in topic CVaR, the analogous baseline entropy term $\log p_{\text{x}|z}(x \mid z)$ depends on z and thus is not a constant with respect to the outer supremum.

In the running toy example (Figure 2), topic CVaR results in robust models that perform relatively well on both news and reviews. The resulting model is a mixture of news and review distribution with equal weights on the two topics.

In summary, topic CVaR contains two improvements over existing DRO approaches: using the latent topic distribution $p_{\text{z}}^{\text{train}}$ to specify the uncertainty set and defining the baselined loss. In the following section, we will describe an algorithm which optimizes this topic CVaR objective.

4 Algorithm

We now operationalize the principles in the previous section, specifying (i) how we choose topics (Section 4.1), (ii) how we estimate the baseline (Section 4.2), and (iii) how to efficiently optimize the robust objective (7) (Section 4.3).

4.1 Identifying Topics

The topic CVaR objective requires topic assignments z for each sentence in order to define the uncertainty set $\mathcal{P}$. Since the topics determine the set of $p_{\text{x}}^{\text{test}}$ distribution for which the model performs well, we seek topics whose subpopulation shifts capture realistic potential test settings.

We use latent Dirichlet allocation (LDA) (Blei et al., 2003) to cluster the sentences into latent topics. LDA assigns each word in a sentence to a topic, and we assign each sentence to the topic with highest total posterior probability.

4.2 Estimating Baselined Losses

Recall that topic CVaR uses KL-divergence as the loss term (Eq. (9)),

$$\text{KL}\big(p_{\text{x}|z} \,\big\|\, p_\theta\big) := \mathbb{E}_{p_{\text{x}|z}}[\log p_{\text{x}|z}(x \mid z)]$$
$$- \mathbb{E}_{p_{\text{x}|z}}[\log p_\theta(x)].$$

While we can estimate the log loss term $\mathbb{E}[\log p_\theta(x)]$ from samples, the entropy term $H(X \mid Z = z) := \mathbb{E}_{p_{\text{x}|z}}[-\log p_{\text{x}|z}(x \mid z)]$ is not something we can easily estimate.

We thus propose to estimate the entropies $H(X \mid Z = z)$ by fitting a *baseline model* p_β for each topic, and computing $H_\beta(X \mid Z = z) := \mathbb{E}_{p_{\text{x}|z}}[-\log p_\beta(x \mid z)]$.[2] In practice, we use a bigram model, which was fast enough to scale and worked sufficiently well in experiments.

4.3 Online Optimization of topic CVaR

No scalable, online algorithm exists for optimizing the topic CVaR objective. Many DRO problems admit efficient batch optimization procedures based on Lagrangian duality (Duchi et al., 2016). However, this approach fails for topic CVaR, since the dual form requires exact computations rather than stochastic estimates of $\mathbb{E}_{p_{\text{x}|z}}[-\log p_\theta(x)]$. Online algorithms for DRO exist (Namkoong and Duchi, 2016), but do not handle the nested maximization-expectation structure arising in topic CVaR (Eq. (7)).

Because of this, we develop an online optimization procedure for topic CVaR compatible with stochastic gradient descent methods. The topic CVaR problem is a two-player minimax game between the model parameter θ and the potential test distribution p_{z}. Intuitively, p_{z} attempts to be the worst-case distribution and maximize the robust objective, while θ attempts to minimize the robust objective. The precise two-player minimax game is

$$\inf_\theta \sup_{p_{\text{z}} \in \mathcal{P}_z^\alpha} \mathbb{E}_{z \sim p_{\text{z}}}\left[L(z; \theta)\right], \qquad (10)$$

where the expected loss for each z (inner expectation) is $L(z; \theta) := \mathbb{E}_{x \sim p_{\text{x}|z}}[\ell(x; \theta)]$.

In the above two-player game, the game proceeds in multiple rounds $t = 1, 2, \ldots$. At each round, the players select $p_{\text{z}}^{(t)}$ and $\theta^{(t)}$. It is standard to interleave parameter updates between the two players in minimax optimization, and we describe the precise update rules in subsequent paragraphs. To carry out these updates, we keep track of an empirical estimate of the probability $p_{\text{z}}^{\text{train}}(z)$ at each iteration t, which we refer to as $\hat{p}_{\text{z}}^{\text{train}(t)}(z)$. We also keep track of the historical average of losses incurred for each topic so far, up to the current round t, which we call $\hat{L}^{(t)}(z; \theta^{(1:t)})$. Con-

---

[2]H_β yields accurate solutions to the topic CVaR problem as long as they capture the entropy up to a constant (i.e. $H_\beta(X \mid Z = z) \approx H(X \mid Z = z) + c$)

cretely, $\hat{L}^{(t)}(z; \theta^{(1:t)})$ is computed as an average of $\{\ell(x^{(t')}; \theta^{(t')}) : t' \in [t], z^{(t')} = z\}$.

At each iteration t, p_z is updated by selecting an optimal value with respect to historical losses up to the current iteration, loosely inspired by the "Be The Leader" algorithm. This results in the following update rule to p_z,

$$p_z^{(t)} = \underset{p_z \in \mathcal{P}_z^\alpha}{\arg\max} \, \mathbb{E}_{z \sim p_z} \left[\hat{L}^{(t)}(z; \theta^{(1:t)}) \right]. \quad (11)$$

The above argmax can computed efficiently by ordering topics in the order of decreasing average loss, and assigning each topic either $\frac{\hat{p}_z^{\text{train}}(z)}{\alpha}$ or the probability left to be assigned, whichever is lower.[3]

We update θ with online gradient descent,

$$\theta^{(t)} = \theta^{(t-1)} - \epsilon \frac{p_z^{(t)}(z^{(t)})}{\hat{p}_z^{\text{train}(t)}(z^{(t)})} \nabla \ell(x^{(t)}; \theta^{(t-1)}),$$

where ϵ is the learning rate.

To give intuition for the two updates, first note that $\frac{p_z^{(t)}(z)}{p_z^{\text{train}(t)}(z)} = \frac{1}{\alpha}$ on approximately α fraction of the data and this ratio acts as an indicator function which determines if an example is part of the worst-case set or not. If it is, we update the model and otherwise we ignore it.

5 Experiments

We demonstrate that topic CVaR improves maximum likelihood language models when $p_x^{\text{train}} \neq p_x^{\text{test}}$. Section 5.1 outlines the experimental setup while Section 5.2 shows the robustness improvements and analysis of topic CVaR.

5.1 Evaluation Details

Datasets. We use the following three corpora: the Yelp review corpus (YELP, (2017)), One Billion Word benchmark corpus (ONEBWORD), and the TripAdvisor Annotated Dataset (TRIPADV, Marcheggiani et al. (2014)).

We preprocess the corpora using SPACY (Honnibal and Johnson (2015)) by removing sentences with fewer than 10 characters, segmenting sentences, tagging named-entities, and replacing each entity with its corresponding OntoNotes tag.

---

[3]For example with $\alpha = 0.2$, $\hat{L}^{(t)} = [40, 30, 60]$, and $\hat{p}_z^{\text{train}} = [0.2, 0.8, 0.1]$, then $p_z^{(t)} = [0.5, 0, 0.5]$.

Vocabulary. Our experiments will evaluate models using perplexity, which depends on the choice of vocabulary. To make perplexity comparable for models trained on different datasets, we use a single, fixed vocabulary formed by combining the most frequently occurring $10,000$ words in each corpus. All words in the mixtures which are not in the vocabulary ($1-3\%$ in our experiments) are replaced with a special *unk* token.

Clustering. To cluster sentences in the training set, we ran LightLDA (Yuan et al. (2015)) for 100 iterations with prior hyperparameters $\alpha = 0.1$, $\beta = 1.0$ and 2 Metropolis-Hastings steps. We set the model to find 10 topics, as this resulted in stable clusters consisting of semantically similar sentences.

Models. Our models are Transformer (Vaswani et al., 2017) based language models trained using the FAIRSEQ sequence-to-sequence toolkit (2017). We use the same model architecture, optimizers, and hyperparameters for both MLE and CVaR. For both models, we use Nesterov's accelerated gradient descent, a fixed learning rate of 0.01, minibatch size of 500 sentences, and 30k minibatches (corresponding to 100 epochs on the YELP corpus). These values were derived by tuning a MLE model trained on the YELP data and tested on the YELP dev set.

Hyperparameters Topic CVaR can be unstable at small α values due to the fact that we are optimizing for worst-case errors. Because of this, we make three small but important modifications to the algorithm. (i) We use $\alpha = 0.2$ to estimate models for $\alpha^* < 0.2$, as small αs can cause gradients to become unstable; (ii) we set a minimum $p_z(z)/p_z^{\text{train}}(z)$ value of 0.1; and (iii) we compute historical losses using exponentially weighted moving averages. With these modifications, the model reliably converges to similar validation losses.

5.2 Language Model Robustness

We seek to assess the performance of MLE and CVaR models under subpopulation shift. In order to do this, we train language models on various mixtures of YELP and ONEBWORD corpora and evaluate the models on a held-out set of YELP sentences.

We will construct a training corpus, whose distribution α^*-*covers* the test distribution (i.e. α^*

fraction of the training distribution corresponds to the Yelp distribution). In this case, we expect that topic CVaR with $\alpha = \alpha^*$ to perform well since the test set exactly fulfills the subpopulation shift assumption.

To form a training corpus, whose distribution α^*-covers the YELP distribution, we mix a fixed set of 500,000 sentences from YELP training subset with $500,000(1 - \alpha^*)/\alpha^*$ sentences from ONEBWORD. This results in a dataset where α^* of the training data comes from YELP. The test corpus is composed of sentences from the YELP test subset, with no sentence overlap with the training corpora. Since the absolute number of YELP samples in the training corpora remains constant across different values of α^*, we expect that a model which is robust to added nuisance data will perform equally well on a YELP-only test set, even as the mixture proportion of ONEBWORD samples in the training corpus increases.

Oracle model. We estimate the *oracle* performance of a robust language model as running topic CVaR where the topic $z = \{$YELP, ONEBWORD$\}$ and the topic assignments use the ground truth corpus identities rather than a clustering algorithm. In this case, when $\alpha^* = \alpha$ we are directly minimizing the worst-case baselined test loss over both YELP and ONEBWORD.
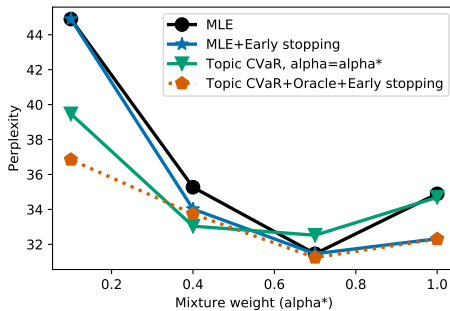

Figure 3. Topic CVaR (green) provides substantial improvements in perplexity compared to MLE (black and blue) as the amount of train-test mismatch increases ($\alpha^* \to 0$). This performance is close to the oracle performance which uses ground truth corpus labels and early stopping (orange).

Topic CVaR improves robustness over MLE. Using the YELP-ONEBWORD mixtures, we evaluate the robustness of topic CVaR and MLE to added nuisance data. We find that with no nuisance data, the MLE model matches the topic CVaR model (Figure 3 $\alpha^* = 1.0$). As we add data from ONEBWORD and α^* decreases to 0.7,

we find some *positive transfer* effects where the increased data from the ONEBWORD corpus improves the performance on Yelp. However, as the fraction of nuisance data grows further and α^* drops below 0.4 the MLE models suffer large increases in perplexity, incurring up to 10 additional points of perplexity. Early stopping according to validation perplexity on YELP does not improve this substantially beyond the basic MLE model (blue star). On the other hand, applying topic CVaR with $\alpha^* = \alpha$ provides substantial boosts to language model performance for small α^*, with nearly no loss of performance for large α^* (green triangle). Finally, we find that the topic CVaR method we propose is close to the best possible *oracle* performance.
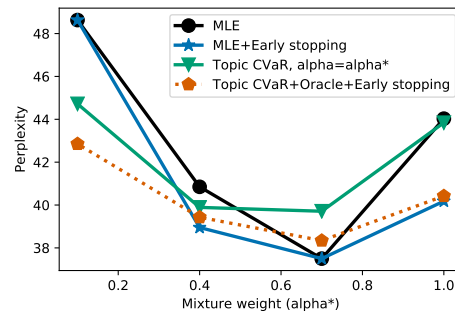

Figure 4. The robustness improvements from topic CVaR (black vs green and orange) apply even when the test set (TRIPADV reviews) is not a subpopulation shift from the training set (YELP and ONEBWORD).

Topic CVaR robustness beyond subpopulation shift. The prior YELP-ONEBWORD experiment showed that topic CVaR is more robust than MLE under subpopulation shift.

We now explore the more realistic setting in which the test distribution is not a subpopulation shift, but merely "similar" to the training distribution. We do this by testing the same model on the TRIPADV hotel review corpus. The hotel and restaurant review distributions are similar (i.e. they both frequently mention service) but differ in that hotels reviews often mention the location and room, while restaurant reviews often mention food items.

We find a similar result consistent with the earlier subpopulation shift experiment (Figure 4). The MLE model performance degrades rapidly between $\alpha^* = 0.7$ and 0.1, while topic CVaR substantially reduces this degradation. This suggests that topic CVaR models provide robustness bene-

fits in real-world settings where the topic overlaps are not exact, and the subpopulation shift assumption no longer holds.

Ablations. Topic CVaR extends the standard CVaR objective in two ways: the use of topics and the use of a baseline. We investigate the effect of these choices via an ablation experiment. Removing the topic structure results in dramatic loss of performance for our models: the perplexity exceeds 80 with $\alpha = 0.2$ for all α^*. This is because the worst case group can consist of solely of disfluent sentences that do not match any real test distribution. If we remove the baseline, the resulting model is not completely degenerate, but it is not as robust as α^* decreases (Figure 5, teal). This is because ONEBWORD is a higher entropy corpus than YELP, and forcing the model to achieve equal *absolute* losses causes the model to focus nearly entirely on ONEBWORD, resulting in low YELP performance.
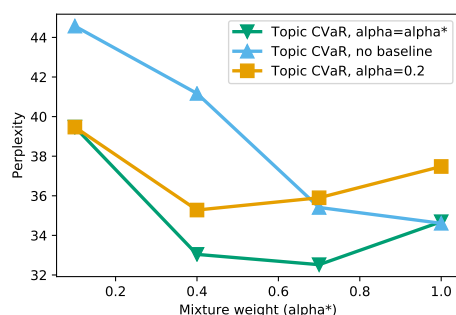

Figure 5. The robustness of topic CVaR degrades when the baseline is removed (teal), but is resistant to being over-conservative in choosing α (yellow).

Choice of α. Since the true train-test overlap α^* is not always known, we cannot always set our hyperparameter α equal to α^*. We find that selecting suboptimal values of α degrades perplexity between 2–3 points depending on α^*. Figure 5 shows that setting α to the most conservative choice of 0.2 outperforms MLE on small α^* while incurring only 2 points of perplexity loss over MLE at $\alpha^* = 1.0$. Figure 6 further demonstrates that when $\alpha^* = 0.1$, any choice of α outperforms MLE, and incorrectly selecting α seems to incur a linear penalty in perplexity.

Error analysis and examples. Evaluating both models trained with $\alpha^* = 0.1$ on both the YELP and ONEBWORD test sets, we find that topic CVaR assigns higher probabilities (and therefore
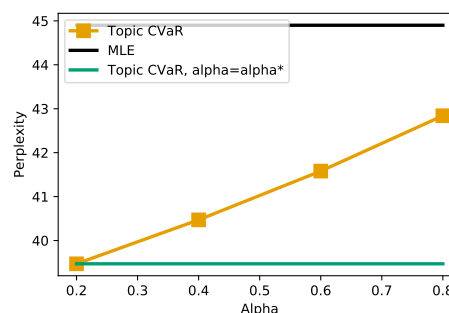

Figure 6. Topic CVaR outperforms MLE in the $p_\mathrm{x}^{\mathrm{train}} \neq p_\mathrm{x}^{\mathrm{test}}$ setting ($\alpha^* = 0.1$) for any small α (x-axis). The performance degradation is linear, implying topic CVaR is robust to small errors in the choice of α.

incurs lower losses) on sentences from Yelp (Figure 7, top right). We also see that MLE does particularly well on low loss examples (bottom left) while topic CVaR does well on high-loss ones (top right) as we might expect from optimizing the worst-case losses.

Examining examples from the YELP test set (Table 1), we identify examples which have substantially higher probabilities under MLE than topic CVaR (left column) and vice versa (right column). These examples show that topic CVaR performs well by assigning high probabilities to stereotypical YELP sentences that discuss food and service, while MLE performs better on sentences about accidents and locations. These examples are consistent with the observation that topic CVaR assigns higher probabilities to typical YELP sentences and thus has lower perplexity, while the MLE model has high perplexity since it assigns probabilities to YELP sentences primarily based on their similarity to examples from ONEB-WORD.
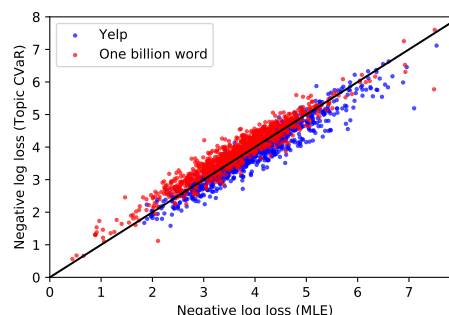

Figure 7. Log losses for sentences (points) from YELP (blue) and ONEBWORD (red) under topic CVaR (y-axis) and MLE (x-axis). Topic CVaR performs well on YELP and infrequent sentences (top right). MLE performs better on frequent sentences from ONEBWORD (bottom left).

$p_{\mathbf{MLE}} > p_{\mathbf{CVaR}}$	$p_{\mathbf{CVaR}} > p_{\mathbf{MLE}}$
my girlfriend had an awful accident that hurt her leg & ankle which resulted in a fire and rescue ride	huge servings, so plenty for leftovers.
the address [PERSON] has listed is their old address	it tastes the way food should taste!
wonderful location in a up and coming part of [GPE].	every single person we spoke to on staff was absolutely incredible.

Table 1. Examples from the YELP corpus for which MLE outperforms topic CVaR (left column) and vice versa. Brackets indicate ONTONOTES named-entity tags. The examples preferred by topic CVaR are stereotypical Yelp sentences, while those preferred by MLE refer to locations and accidents.

6 Related Work

Domain Adaptation: In the case of known source (train) and target (test) domains, there exist a variety of techniques to learn robust models (Shimodaira, 2000; Quiñonero-Candela et al., 2009; Daume III, 2007; Ben-David et al., 2010; Blitzer et al., 2011; Pryzant et al., 2017) or domain-invariant features (Ganin and Lempitsky, 2015; Tzeng et al., 2014). However, such methods require accurate domain membership annotations.

In the absence of domain membership annotations, prior multi-source domain adaptation (Mansour et al., 2009) approaches propose the use of clustering to identify candidate domains. For instance, Hoffman et al. (2012) and Xiong et al. (2014) discover latent domains in classification by clustering data using class labels. Gong et al. (2013) extend this work by identifying subsets which are distinct and learnable. More recent work consider errors in estimating the target domain (Hoffman et al., 2018) and derive learning bounds with respect to such errors. While these approaches make use of cluster and topic structures as prior, they still require some knowledge of the target distribution and train a model tailored to the target distribution. Instead, we assume no knowledge on the target distribution and train a single model by considering the worst case.

In *conditional* settings such as machine translation, prior works connect topic modeling and domain adaptation (Hu et al., 2014; Eidelman et al., 2012). However, unlike our work, these approaches use topics at *test time* by inferring the domain from the input variable x. In language modeling, we have no inputs and thus must find models robust to unknown domain shifts at test time. In addition, it can be difficult to infer the test distribution as the distribution can rapidly change across users and time.

Distributional Robustness: Our approach is based upon existing work in the distributionally robust optimization (DRO) literature. Optimizing on a ball of distributions around the empirical distribution has been considered in prior work (Ben-Tal et al., 2013; Namkoong and Duchi, 2017; Duchi and Namkoong, 2016; Sinha et al., 2018). Using DRO to minimize losses over subpopulations was proposed earlier in Hashimoto et al. (2018) and Duchi and Namkoong (2018), and Hu et al. (2018) proposed incorporating problem structure via class labels. Our work derives an efficient optimization procedure for DRO with topic-based uncertainty sets, and demonstrates that naively applying DRO to log losses fails to provide robustness due to the lack of baselining.

7 Discussion

In this work, we show that the performance of language models degrade as the amount of text from outside the test distribution grows. We hypothesize that this problem arises from the tendency of MLE to optimize for common sentences in the corpus, and we propose a solution based on distributionally robust optimization.

Empirically, we demonstrate that the DRO-based topic CVaR is more robust than MLE to subpopulation shifts and similar shifts. While this work focuses on DRO for language modeling, train-test mismatches under subpopulation shifts are more broadly applicable to any task where there are trade-offs between potential test distributions, and potential test distributions can be described with topics. Our work shows that topics are an effective way to encode prior information about test distributions, and baselines can properly normalize for the difficulty across these topics.

Reproducibility. Code and data is available on CodaLab: `https://worksheets.codalab.org/worksheets/0xf8122ebd24e94209a2a1764007509098`.

References

S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. W. Vaughan. 2010. A theory of learning from different domains. *Machine Learning*, 79(1):151–175.

A. Ben-Tal, D. den Hertog, A. D. Waegenaere, B. Melenberg, and G. Rennen. 2013. Robust solutions of optimization problems affected by uncertain probabilities. *Management Science*, 59:341–357.

D. Blei, A. Ng, and M. I. Jordan. 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research (JMLR)*, 3:993–1022.

J. Blitzer, S. Kakade, and D. P. Foster. 2011. Domain adaptation with coupled subspaces. In *Artificial Intelligence and Statistics (AISTATS)*, pages 173–181.

C. Chelba, T. Mikolov, M. Schuster, Q. Ge, T. Brants, P. Koehn, and T. Robinson. 2013. One billion word benchmark for measuring progress in statistical language modeling. *arXiv preprint arXiv:1312.3005*.

A. M. Dai and Q. V. Le. 2015. Semi-supervised sequence learning. In *Advances in Neural Information Processing Systems (NeurIPS)*.

H. Daume III. 2007. Frustratingly easy domain adaptation. In *Association for Computational Linguistics (ACL)*.

J. Devlin, M. Chang, K. Lee, and K. Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

J. Duchi, P. Glynn, and H. Namkoong. 2016. Statistics of robust optimization: A generalized empirical likelihood approach. *arXiv*.

J. Duchi and H. Namkoong. 2016. Variance-based regularization with convex objectives. *arXiv preprint arXiv:1610.02581*.

J. Duchi and H. Namkoong. 2018. Learning models with uniform performance via distributionally robust optimization. *arXiv preprint arXiv:1810.08750*.

V. Eidelman, J. Boyd-Graber, and P. Resnik. 2012. Topic models for dynamic translation model adaptation. In *Association for Computational Linguistics (ACL)*, pages 115–119.

Y. Ganin and V. Lempitsky. 2015. Unsupervised domain adaptation by backpropagation. In *International Conference on Machine Learning (ICML)*, pages 1180–1189.

J. Gehring, M. Auli, D. Grangier, D. Yarats, and Y. N. Dauphin. 2017. Convolutional sequence to sequence learning. *arXiv preprint arXiv:1705.03122*.

B. Gong, K. Grauman, and F. Sha. 2013. Reshaping visual datasets for domain adaptation. In *Advances in Neural Information Processing Systems (NeurIPS)*.

T. B. Hashimoto, M. Srivastava, H. Namkoong, and P. Liang. 2018. Fairness without demographics in repeated loss minimization. In *International Conference on Machine Learning (ICML)*.

J. Hoffman, B. Kulis, T. Darrell, and K. Saenko. 2012. Discovering latent domains for multisource domain adaptation. In *European Conference on Computer Vision (ECCV)*, pages 702–715.

J. Hoffman, M. Mohri, and N. Zhang. 2018. Algorithms and theory for multiple-source adaptation. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 8256–8266.

M. Honnibal and M. Johnson. 2015. An improved non-monotonic transition system for dependency parsing. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1373–1378.

W. Hu, G. Niu, I. Sato, and M. Sugiyama. 2018. Does distributionally robust supervised learning give robust classifiers? In *International Conference on Machine Learning (ICML)*.

Y. Hu, K. Zhai, V. Eidelman, and J. Boyd-Graber. 2014. Polylingual tree-based topic models for translation domain adaptation. In *Association for Computational Linguistics (ACL)*, pages 1166–1176.

Y. Mansour, M. Mohri, and A. Rostamizadeh. 2009. Domain adaptation with multiple sources. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 1041–1048.

D. Marcheggiani, O. Täckström, A. Esuli, and F. Sebastiani. 2014. Hierarchical multi-label conditional random fields for aspect-oriented opinion mining. In *ECIR*.

B. McCann, J. Bradbury, C. Xiong, and R. Socher. 2017. Learned in translation: Contextualized word vectors. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 6297–6308.

R. Nallapati, B. Zhou, C. Gulcehre, B. Xiang, et al. 2016. Abstractive text summarization using sequence-to-sequence rnns and beyond. *arXiv preprint arXiv:1602.06023*.

H. Namkoong and J. Duchi. 2016. Stochastic gradient methods for distributionally robust optimization with f-divergences. In *Advances in Neural Information Processing Systems (NeurIPS)*.

H. Namkoong and J. Duchi. 2017. Variance regularization with convex objectives. In *Advances in Neural Information Processing Systems (NeurIPS)*.

M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer. 2018. Deep contextualized word representations. In *North American Association for Computational Linguistics (NAACL)*.

R. Pryzant, D. Britz, and Q. V. Le. 2017. Effective domain mixing for neural machine translation. In *Proceedings of the Second Conference on Machine Translation*, pages 118–126.

J. Quiñonero-Candela, M. Sugiyama, A. Schwaighofer, and N. D. Lawrence. 2009. *Dataset shift in machine learning*. The MIT Press.

A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever. 2018. Improving language understanding by generative pre-training. Technical report, OpenAI.

R. T. Rockafellar and S. Uryasev. 2000. Optimization of conditional value-at-risk. *Journal of Risk*, 2:21–41.

H. Shimodaira. 2000. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, 90:227–244.

A. Sinha, H. Namkoong, and J. Duchi. 2018. Certifiable distributional robustness with principled adversarial training. In *International Conference on Learning Representations (ICLR)*.

A. Sordoni, M. Galley, M. Auli, C. Brockett, Y. Ji, M. Mitchell, J. Nie, J. Gao, and B. Dolan. 2015. A neural network approach to context-sensitive generation of conversational responses. In *North American Association for Computational Linguistics (NAACL)*.

E. Tzeng, J. Hoffman, N. Zhang, K. Saenko, and T. Darrell. 2014. Deep domain confusion: Maximizing for domain invariance. *arXiv preprint arXiv:1412.3474*.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. 2017. Attention is all you need. *arXiv preprint arXiv:1706.03762*.

A. Vaswani, Y. Zhao, V. Fossum, and D. Chiang. 2013. Decoding with large-scale neural language models improves translation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1387–1392.

C. Xiong, S. McCloskey, S. Hsieh, and J. J. Corso. 2014. Latent domains modeling for visual domain adaptation. In *Association for the Advancement of Artificial Intelligence (AAAI)*.

Yelp. 2017. Yelp Dataset Challenge, Round 8. https://www.yelp.com/dataset_challenge.

J. Yuan, F. Gao, Q. Ho, W. Dai, J. Wei, X. Zheng, E. P. Xing, T. Liu, and W. Ma. 2015. Lightlda: Big topic models on modest compute clusters. In *World Wide Web (WWW)*.