On the KL-Divergence based Robust Satisficing Model

Haojie Yan

hjyan21@m.fudan.edu.cn

University of Fudan

School of Information Science and Technology

June 9, 2024

Presentation Overview

Stochastic Optimization

We start from considering a stochastic optimization problem as follows:

$$\min_{\boldsymbol{\theta}\in\Theta} \mathbb{E}_{\mathbb{P}^*}[l(\boldsymbol{\theta}, \tilde{\boldsymbol{z}})], \tag{1}$$

where θ is the decision variable with feasible region Θ, $\tilde{\boldsymbol{z}}$ represents random variables satisfying joint distribution $\mathbb{P}^*$. $l(\boldsymbol{\theta}, \tilde{\boldsymbol{z}})$ is loss function.

- Pros: In many cases, the expected value is a good measure of performance.
- Cons: One has to know the exact distribution of $\tilde{\boldsymbol{z}}$ to perform the stochastic optimization!

Empirical Risk Minimization

In practice, although the exact distribution of the random variables can not be known in advance. People usually may have some certain observed samples or training data. Estimating the exact distribution $\mathbb{P}^*$ by on-hand samples is a natural method. One of the most classical estimations is empirical distribution.

- With N historical observation $\{\hat{\boldsymbol{z}}_i\}_{i\in[N]}$.
- Empirical Distribution $\hat{\mathbb{P}}(\tilde{\boldsymbol{z}} = \hat{\boldsymbol{z}}_i) = 1/N$.

$$\min_{\boldsymbol{\theta}\in\Theta} \mathbb{E}_{\hat{\mathbb{P}}}[l(\boldsymbol{\theta}, \tilde{\boldsymbol{z}})] = \min_{\boldsymbol{\theta}\in\Theta} \frac{1}{N}\sum_{i=1}^{N} l(\boldsymbol{\theta}, \hat{\boldsymbol{z}}_i). \tag{2}$$

Motivation of Distributionally Robust Method

1. **Optimizer's Curse**: The solution of (2) performs poorly out of sample test when historical data is not sufficient.
2. **Dirty Data**: The data is affected by noise or contains missing values.
3. **Distributional Shift**: The test distribution may be different from the train distribution *i.e.* dynamics, domain adaptation.

Distributionally Robust Approach

A solution to address the optimizer's curse is taking Distributionally Robust Optimization (DRO). DRO considers a distribution ambiguity set and optimize the worst case expected performance.

$$\min_{\boldsymbol{\theta}\in\Theta} \max_{\mathbb{P}\in\mathcal{B}(r)} \mathbb{E}_{\mathbb{P}}[l(\boldsymbol{\theta}, \tilde{\boldsymbol{z}})] \quad \text{where } \mathcal{B}(r) \triangleq \{\mathbb{P} \in \mathcal{P}(\Omega) : D(\mathbb{P}\|\hat{\mathbb{P}}) \leq r\}. \tag{3}$$

- D is probability distance which measures the difference between $\mathbb{P}$ and $\hat{\mathbb{P}}$ *i.e.* Wasserstein distance, KL divergence, ϕ divergence.
- DRO provides upper bound for model's performance on distributions inside $\mathcal{B}(r)$.
- There is a high probability that the true distribution is within the distribution ambiguity set $\mathcal{B}(r)$.

Motivation for Robust Satisficing

Shortage of DRO

1. No theoretical guarantee for model performance outside $\mathcal{B}(r)$.
2. Ambiguity set radius r is abstract for determining for decision maker. Decision maker is usually familiar with specific performance.
3. The DRO bound is loose when distribution shift is small.

Recently, a novel development in optimization under uncertainty is the Robust Satisficing (RS) framework, which can provide performance guarantees under all possible only distributions rather than distributions in $\mathcal{B}(r)$.

$$\begin{aligned} \min_{\lambda \geq 0, \boldsymbol{\theta} \in \Theta} \quad & \lambda \\ \text{s.t.} \quad & \mathbb{E}_{\mathbb{P}}[l(\boldsymbol{\theta}, \tilde{\boldsymbol{z}})] \leq \tau + \lambda D(\mathbb{P} \| \hat{\mathbb{P}}) \quad \forall \mathbb{P} \in \mathcal{P}(\Omega), \end{aligned} \tag{4}$$

where τ is a target value which is an acceptable performance value for decision maker.

Investigation Gap

1. Existing research on RS usually adopts Wasserstein distance. However, Wasserstein distance fails to capture non-geometric distributional shift, *i.e.* kernel distribution shift in MDP, label distribution shift and domain adaptation. While KL divergence is suitable for accounting such shift.
2. KL divergence based RS model in machine learning has not been well investigated.
3. When the loss function is complex function such as neural network, the solving method has not been investigated.

Problem Formulation

KL Divergence-based Robust Satisficing Model (KL-RS) can be formulated as the following form:

$$\begin{aligned} \min_{\lambda \geq 0, \boldsymbol{\theta} \in \Theta} \quad & \lambda \\ \text{s.t.} \quad & \mathbb{E}_{\mathbb{P}}[l(\boldsymbol{\theta}, \tilde{\boldsymbol{z}})] \leq \tau + \lambda D_{KL}(\mathbb{P} \| \hat{\mathbb{P}}) \quad \forall \mathbb{P} \ll \hat{\mathbb{P}}. \end{aligned} \tag{5}$$

where $D_{KL}(\mathbb{P} \| \hat{\mathbb{P}})$ is Kullbak-Leibler divergence defined as follows:

$$D_{KL}(\mathbb{Q} \| \mathbb{P}) = \mathbb{E}_{\mathbb{Q}}\left[\log\left(\frac{d\mathbb{Q}}{d\mathbb{P}}\right)\right] \quad \text{if} \quad \mathbb{Q} \ll \mathbb{P}, \tag{6}$$

where $\ll$ denotes that $\mathbb{Q}$ is absolute continuous with $\hat{\mathbb{P}}$.

Tractable Reformulation

Directly optimize (5) is difficult because we need to optimize $\mathbb{P}$ and $\boldsymbol{\theta}$ simultaneously. Some duality properties can be adopted to simplify our formulation.

Theorem

The KL-RS model (5) *is equivalent to*

$$\begin{aligned} \min_{\lambda \geq 0, \boldsymbol{\theta} \in \Theta} \quad & \lambda \\ s.t. \quad & \hat{R}(\boldsymbol{\theta}, \lambda) \leq \tau, \end{aligned} \tag{7}$$

with $\hat{R}(\boldsymbol{\theta}, \lambda) \triangleq \lambda \log \left(\mathbb{E}_{\hat{\mathbb{P}}} \left[\exp \left(l(\boldsymbol{\theta}, \tilde{\boldsymbol{z}}) / \lambda \right) \right] \right)$.

Analytical Interpretation

The specific KL-RS model has its own unique properties that are worth analyzing.
As a concrete example, let us consider a linear loss function $l(\boldsymbol{\theta}, \tilde{\boldsymbol{z}}) = \boldsymbol{\theta}^\top \tilde{\boldsymbol{z}}$ and $\tilde{\boldsymbol{z}} \sim N(\mu, \Sigma)$.

$$\hat{R}(\boldsymbol{\theta}, \lambda) = \lambda \log \left(\mathbb{E}_{\hat{\mathbb{P}}} \left[\exp \left(l(\boldsymbol{\theta}, \tilde{\boldsymbol{z}}) / \lambda \right) \right] \right) = \boldsymbol{\theta}^\top \mu + \frac{\boldsymbol{\theta}^\top \Sigma \boldsymbol{\theta}}{2\lambda} \leq \tau.$$

KL-RS model robustify the solution by increasing the weight of the variance while ensuring the weighted mean and varaince is bouned below τ.

Proposition

If λ is large enough, we have $\hat{R}(\boldsymbol{\theta}, \lambda) = \mathbb{E}_{\hat{\mathbb{P}}}[l(\boldsymbol{\theta}, \tilde{\boldsymbol{z}})] + \frac{1}{2\lambda} \mathbb{V}_{\hat{\mathbb{P}}}[l(\boldsymbol{\theta}, \tilde{\boldsymbol{z}})] + o(\frac{1}{\lambda^2})$, where $\mathbb{V}_{\hat{\mathbb{P}}}[l(\boldsymbol{\theta}, \tilde{\boldsymbol{z}})]$ is the variance of $l(\boldsymbol{\theta}, \tilde{\boldsymbol{z}})$.

Prioritization On Large Losses

With a little transformation of original KL-RS model (5), we have the following formulation:

$$\begin{aligned} \min_{\lambda \geq 0, \boldsymbol{\theta} \in \Theta} \quad & \lambda \\ \text{s.t.} \quad & \sup_{\mathbb{P} \ll \hat{\mathbb{P}}} \left\{ \mathbb{E}_{\mathbb{P}}[l(\boldsymbol{\theta}, \tilde{\boldsymbol{z}})] - \lambda D_{KL}(\mathbb{P} \| \hat{\mathbb{P}}) \right\} \leq \tau. \end{aligned} \tag{8}$$

For a given $\boldsymbol{\theta}$ and λ, we let $\mathbb{P}_0^*$ to denote the worst case distribution,

$$\mathbb{P}_0^*(\tilde{\boldsymbol{z}}_0^* = \hat{\boldsymbol{z}}_i) = \frac{\exp(l(\boldsymbol{\theta}, \hat{\boldsymbol{z}}_i)/\lambda)}{N \cdot \mathbb{E}_{\hat{\mathbb{P}}}[\exp(l(\boldsymbol{\theta}, \tilde{\boldsymbol{z}})/\lambda)]}.$$

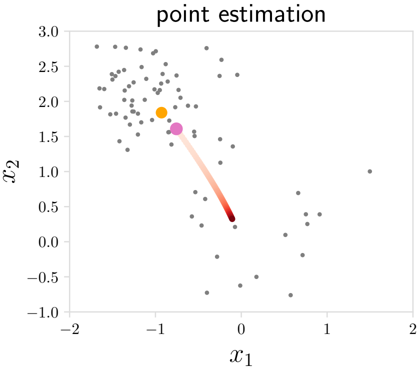
KL-RS model magnify the impact of samples with large losses.

Two extreme cases

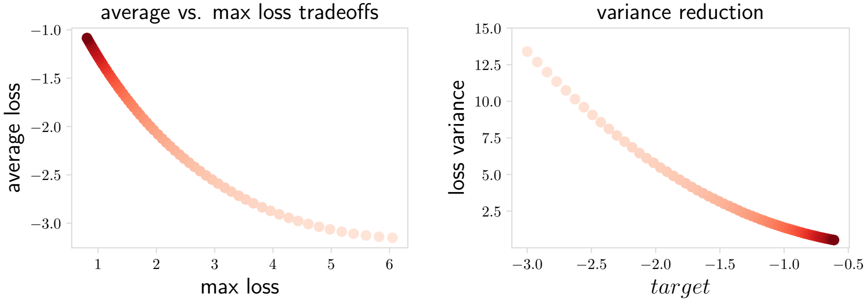
$$\lim_{\lambda \to 0^+} \hat{R}(\boldsymbol{\theta}, \lambda) = \max_{i \in [N]} l(\hat{\boldsymbol{z}}_i, \boldsymbol{\theta}), \quad \lim_{\lambda \to +\infty} \hat{R}(\boldsymbol{\theta}, \lambda) = \mathbb{E}_{\hat{\mathbb{P}}}[l(\tilde{\boldsymbol{z}}, \boldsymbol{\theta})]$$

Point Estimation

- 100 two-dimensional points generated from two normal distributions.
- $l(\boldsymbol{\theta}, \tilde{\boldsymbol{z}}) = \frac{1}{2}(\boldsymbol{\theta} - \tilde{\boldsymbol{z}})^2$.
- Estimation resulted by larger τ with deeper color.



Experiment Result



(a) (b)

Figure: Performance evaluations across varied τ

Experiment Result

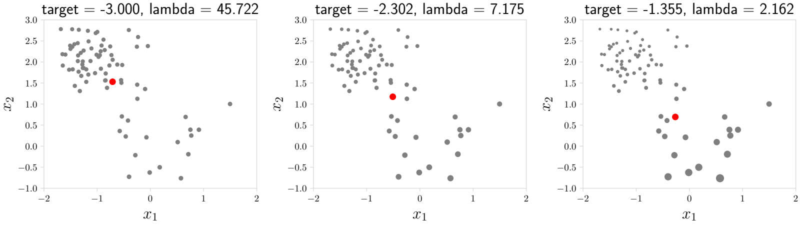


Figure: Weights attributed to samples

Tail Probability Guarantee

We begin the discussion about the properties of KL-RS solution with its tail performance on empirical distribution.

Proposition

Given a non-negative number α, we have

$$\hat{\mathbb{P}}(l(\boldsymbol{\theta}, \tilde{\boldsymbol{z}}) \geq \tau + \alpha) \leq \exp(-\alpha/\lambda)$$

for every feasible solution $(\boldsymbol{\theta}, \lambda)$ of the KL-RS model.

- Probability exceeding the tolerance τ decays with exponentially.
- In some sense, this proposition provide analytical interpretation of KL-RS from another perspective, KL-RS concentrates empirical loss from above.

Asymptotic Performance Guarantee

The performance of the solution of KL-RS model on true distribution.
Let $(\boldsymbol{\theta}_N^*, \lambda_N^*)$ denote the optimal solution of KL-RS model. The following inequality holds:

$$\mathbb{E}_{\mathbb{P}}[l(\boldsymbol{\theta}_N^*, \tilde{\boldsymbol{z}})] \le \tau + \lambda_N^* D_{KL}(\mathbb{P} \| \hat{\mathbb{P}}), \quad \mathbb{P} \ll \hat{\mathbb{P}}.$$

For a non-negative r, if we know the probability of event where $D_{KL}(\mathbb{P} \| \hat{\mathbb{P}}) \le r$. Let $\mathbb{P}^N$ denote the distribution which governs the distribution of independent samples $\{\hat{\boldsymbol{z}}_i\}_{i \in [N]}$.
$\mathbb{P}^N$ is the N-fold of Cartesian product of $\mathbb{P}^*$.

Asymptotic Performance Guarantee

Theorem

Suppose that $\mathbb{P}^*$ is a discrete distribution supported by K points. For optimal solution $(\boldsymbol{\theta}_N^*, \lambda_N^*)$ of the KL-RS model and a given non-negative radius $r \geq 0$, we have

$$\mathbb{P}^N \left(\mathbb{E}_{\mathbb{P}^*}[l(\boldsymbol{\theta}_N^*, \tilde{\mathbf{z}})) < \tau + \lambda_N^* r\right) \geq \chi^2_{K-1}(\tilde{y} \leq 2Nr) \quad \text{as} \quad N \to \infty, \tag{9}$$

where we use $\mathbb{P}^N$ to denote the distribution that governs the distribution of independent samples $\{\hat{\mathbf{z}}_i\}_{i\in[N]}$ drawn from $\mathbb{P}^*$, and $\tilde{y} \sim \chi^2_{K-1}$ is a chi-squared distribution with degree of freedom $K-1$.

Finite Sample Guarantee

Theorem

Suppose that $\mathbb{P}^*$ is a discrete distribution supported by K points, and $\hat{\mathbb{P}}^l$ is a Laplace smoothing of $\hat{\mathbb{P}}$ with N samples. For optimal solution $(\boldsymbol{\theta}_N^*, \lambda_N^*)$ of the KL-RS model on $\hat{\mathbb{P}}^l$ and a threshold $\delta > 0$, we have

$$\mathbb{P}^N \left(\mathbb{E}_{\mathbb{P}^*}[l(\boldsymbol{\theta}_N^*, \tilde{\boldsymbol{z}})) < \tau + \lambda_N^* r \right) \geq 1 - \delta$$

The where we use $\mathbb{P}^N$ to denote the distribution that governs the distribution of independent samples $\{\hat{\boldsymbol{z}}_i\}_{i \in [N]}$ drawn from $\mathbb{P}^*$, and
$r = \mathbb{E}_{\mathbb{P}^N}[D_{KL}(\mathbb{P}^* \| \hat{\mathbb{P}}^l)] + \frac{6\sqrt{K \log^5(4K/\delta)} + 311}{N} + \frac{160K}{N^{3/2}}$.

Hierarchical KL-RS

Observation: two layer hierarchical structures is commonly seen across various fields.

- classification task, fair machine learning, agnostic machine learning, invariant risk minimization, contextual stochastic bilevel optimization ...

Introduce an extra random variable $\tilde{\boldsymbol{g}}$ to denote the group information. Use $\mathbb{P}$ to denote the joint distribution $(\tilde{\boldsymbol{z}}, \tilde{\boldsymbol{g}})$, $\mathbb{P}_{\tilde{\boldsymbol{g}}}$ and $\mathbb{P}_{\tilde{\boldsymbol{z}}|\tilde{\boldsymbol{g}}}$ to denote marginal distribution and conditional distribution.

Let w be a non-negative weight hyperparameter. Hierarchical KL-RS can be formulated as the following:

$$\begin{aligned} \min \quad & \lambda_1 + w\lambda_2 \\ \text{s.t.} \quad & \mathbb{E}_{\mathbb{P}}\left[l(\boldsymbol{\theta}, \tilde{\boldsymbol{z}})\right] \leq \tau + \lambda_1 D_{KL}(\mathbb{P}_{\tilde{\boldsymbol{g}}} \| \hat{\mathbb{P}}_{\tilde{\boldsymbol{g}}}) + \lambda_2 \mathbb{E}_{\mathbb{P}_{\tilde{\boldsymbol{g}}}} D_{KL}(\mathbb{P}_{\tilde{\boldsymbol{z}}|\tilde{\boldsymbol{g}}} \| \hat{\mathbb{P}}_{\tilde{\boldsymbol{z}}|\tilde{\boldsymbol{g}}}). \end{aligned} \tag{10}$$

A tractable formulation:

$$\begin{aligned} \min_{\boldsymbol{\theta}\in\Theta, \lambda_1 \geq 0, \lambda_2 \geq 0} \quad & \lambda_1 + w\lambda_2 \\ \text{s.t.} \quad & \hat{R}(\boldsymbol{\theta}, \lambda_1, \lambda_2) \leq \tau. \end{aligned}$$

, $\hat{R}(\boldsymbol{\theta}, \lambda_1, \lambda_2) \triangleq \lambda_1 \log\left(\mathbb{E}_{\hat{\mathbb{P}}_{\tilde{\boldsymbol{g}}}}\left[\exp\left(\lambda_2 \log\left(\mathbb{E}_{\hat{\mathbb{P}}_{\tilde{\boldsymbol{z}}|\tilde{\boldsymbol{g}}}} \exp\left(l(\boldsymbol{\theta}, \tilde{\boldsymbol{z}})/\lambda_2\right)\right)/\lambda_1\right)\right]\right)$

Alternative Optimization

We alternatively optimize λ and $\boldsymbol{\theta}$.

- For a given $\boldsymbol{\theta}$, we can find corresponding optimal λ by bisection search. With tolerance ϵ, bisection algorithm will convergence in $O(\log(1/\epsilon))$ iterations.

Algorithm 2: Solve the KL-RS by bisection method

Initialization : $\underline{\lambda} = 0$, a positive value λ_0, a precision $\epsilon > 0$

while Algorithm 1(λ_0) == False **do**
    $\underline{\lambda} \leftarrow \lambda_0$, $\lambda_0 \leftarrow 2\lambda_0$
end
    $\overline{\lambda} = \lambda_0$
while $\overline{\lambda} - \underline{\lambda} \geq \epsilon$ **do**
    $\lambda_{\text{mid}} = (\overline{\lambda} + \underline{\lambda})/2$
    if Algorithm 1(λ_{mid})==True **then**
        $\overline{\lambda} = \lambda_{\text{mid}}$
    end
    else
        $\underline{\lambda} = \lambda_{\text{mid}}$
    end
end
Output: λ_{mid}

Alternative Optimization

- For a given λ, note that $\hat{R}(\boldsymbol{\theta}, \lambda) \leq \tau \Leftrightarrow \mathbb{E}_{\hat{\mathbb{P}}}[f(\boldsymbol{\theta}, \tilde{\boldsymbol{z}}; \lambda)] \leq 1$ with $f(\boldsymbol{\theta}, \tilde{\boldsymbol{z}}; \lambda) \triangleq \exp\left(\frac{l(\boldsymbol{\theta}, \tilde{\boldsymbol{z}}) - \tau}{\lambda}\right)$. We can perform gradient based algorithm to solve

$$\min_{\boldsymbol{\theta}} \mathbb{E}_{\hat{\mathbb{P}}}[f(\boldsymbol{\theta}, \tilde{\boldsymbol{z}}; \lambda)].$$

  $f(\boldsymbol{\theta}, \tilde{\boldsymbol{z}}; \lambda)$ can inherent some good optimization properties from $l(\boldsymbol{\theta}, \tilde{\boldsymbol{z}})$, *i.e.* convexity, strongly convexity, lipschitz continuous and lipschitz smooth. The convergence rate of gradient based algorithm on $f(\boldsymbol{\theta}, \tilde{\boldsymbol{z}}; \lambda)$ is the same as on $l(\boldsymbol{\theta}, \tilde{\boldsymbol{z}})$.

Solving Hierarchical KL-RS

Similar to solving standard KL-RS but more complicated.

- When $\boldsymbol{\theta}$ is fixed, we can also adopt search based method to solve the optimal (λ_1, λ_2).

Proposition

For a given $\boldsymbol{\theta}$, $\hat{R}(\boldsymbol{\theta}, \lambda_1, \lambda_2)$ is convex with (λ_1, λ_2). Combining golden search and bisection search.

Algorithm 4: Find optimal λ_2 by bisection method (λ_1)

Input: λ_1
Initialization : $\underline{\lambda} = 0$, a positive value λ_0, a precision $\epsilon > 0$
while Algorithm 3(λ_1, λ_0) == False **do**
  $\underline{\lambda} \leftarrow \lambda_0$, $\lambda_0 \leftarrow 2\lambda_0$
end
  $\overline{\lambda} = \lambda_0$
while $\overline{\lambda} - \underline{\lambda} \geq \epsilon$ **do**
  $\lambda_{\text{mid}} = (\overline{\lambda} + \underline{\lambda})/2$
  if Algorithm 3$(\lambda_1, \lambda_{\text{mid}})$==True **then**
    $\overline{\lambda} = \lambda_{\text{mid}}$
  end
  else
    $\underline{\lambda} = \lambda_{\text{mid}}$
  end
end
Output: λ_{mid}

Algorithm 5: Solve the hierarchical KL-RS by golden-ratio search

Initialization : a precision $\epsilon > 0$, $\gamma = 0.382$, $\lambda_l = \lambda_{\min}$, $\lambda_r = \lambda_{\max}$
while $\lambda_r - \lambda_l \geq \epsilon$ **do**
  $\lambda'_l = \lambda_l + \gamma(\lambda_r - \lambda_l)$, $\lambda'_r = \lambda_l + (1-\gamma)(\lambda_r - \lambda_l)$
  $\lambda_l^{(2)}$ =Algorithm 4(λ'_l), $\lambda_r^{(2)}$ =Algorithm 4(λ'_r)
  if $\lambda'_l + w\lambda_l^{(2)} \leq \lambda'_r + w\lambda_r^{(2)}$ **then**
    $\lambda_r = \lambda'_r$
  end
  else
    $\lambda_l = \lambda'_l$
  end
end
Output: λ_{mid}

- For given (λ_1, λ_2), note that

$$
\begin{aligned}
& \hat{R}(\boldsymbol{\theta}, \lambda_1, \lambda_2) \le \tau \\
\iff \quad & \mathbb{E}_{\hat{\mathbb{P}}_{\tilde{\boldsymbol{g}}}}\left[h\left(\mathbb{E}_{\hat{\mathbb{P}}_{\tilde{\boldsymbol{z}}|\tilde{\boldsymbol{g}}}}\left[f(\boldsymbol{\theta}, \tilde{\boldsymbol{z}}; \lambda_2) \right]; \lambda_1, \lambda_2 \right) \right] \le 1,
\end{aligned}
$$

  where $h(x; \lambda_1, \lambda_2) \triangleq x^{\frac{\lambda_2}{\lambda_1}}$.

- Conditional Stochastic Optimization (CSO).

$$
\min_{\boldsymbol{\theta} \in \Theta} \mathbb{E}_{\hat{\mathbb{P}}_{\tilde{\boldsymbol{g}}}}\left[h\left(\mathbb{E}_{\hat{\mathbb{P}}_{\tilde{\boldsymbol{z}}|\tilde{\boldsymbol{g}}}}\left[f(\boldsymbol{\theta}, \tilde{\boldsymbol{z}}; \lambda_2) \right]; \lambda_1, \lambda_2 \right) \right]
$$

  CSO problem is computationally challenging because the estimator of objective and gradient is usually biased. The conditional distribution makes sampling and optimization more complex.

Solving Hierarchical KL-RS

Bilevel Sample Method

1. We iid sample M_1 $\hat{\boldsymbol{g}}_i$ from coniditional distribution $\hat{\mathbb{P}}_{\tilde{\boldsymbol{g}}}$.
2. For each $\hat{\boldsymbol{g}}_i$, we iid sample M_2 $\hat{\boldsymbol{z}}_{i,j}$ from condistional distribution $\hat{\mathbb{P}}_{\tilde{\boldsymbol{z}}|\hat{\boldsymbol{g}}}$

Algorithm 3: Feasibility of the hierarchical KL-RS model (λ_1, λ_2)

Input: λ_1, λ_2
Initialization: group batch size M_1, batch size within group M_2, step size α
while *stopping criteria not reached* **do**
    Sample $\hat{\boldsymbol{g}}_i$ uniformly random from $\tilde{\boldsymbol{g}}$ with batch size M_1;
    for $i = 1$ *to* M_1 **do**
        For each given $\hat{\boldsymbol{g}}_i$, sample $\hat{\boldsymbol{z}}_{i,j}$ uniformly random from $\tilde{\boldsymbol{z}}|\hat{\boldsymbol{g}}_i$ with batch size M_2;
        Construct $\bar{l}_i \triangleq \frac{1}{M_2} \sum_{j \in [M_2]} l(\boldsymbol{\theta}, \hat{\boldsymbol{z}}_{i,j}), \quad \bar{\mathrm{l}}_i \triangleq \frac{1}{M_2} \sum_{j \in [M_2]} \nabla l(\boldsymbol{\theta}, \hat{\boldsymbol{z}}_{i,j})$
    end
    Construct $\nabla F \triangleq \frac{1}{M_1} \sum_{i \in [M_1]} h'(\bar{l}_i; \lambda_1, \lambda_2) \cdot \bar{\mathrm{l}}_i$
    Update $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \alpha \nabla F$
end
Output: Boolean$\left(\mathbb{E}_{\hat{\mathbb{P}}_{\tilde{\boldsymbol{g}}}} \left[h\left(\mathbb{E}_{\hat{\mathbb{P}}_{\tilde{\boldsymbol{z}}|\tilde{\boldsymbol{g}}}} \left[f(\boldsymbol{\theta}, \tilde{\boldsymbol{z}}; \lambda_2) \right]; \lambda_1, \lambda_2 \right) \right] \le 1 \right)$

Label Distribution Shift

The test set label distribution is different from the train set label distribution.

Problem Settings

- Binary Classification Task
- Dataset: HIV-1 dataset, positive samples are much fewer than negative samples. (Imbalanced Dataset)
- The ERM solution performs poorly when the label distribution shift happens.
- Our KL-RS model enjoys profile of performance guarantees with respect to the expected prediction loss under all possible distributions.

Experiment Design

We want to evaluate the model's performance on test distribution with different KL divergence form train set.

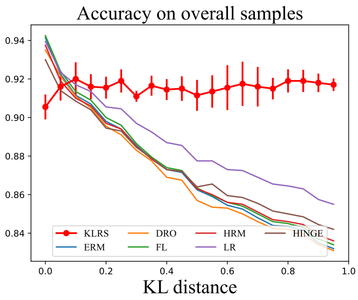
- Evaluation Metric: Accuracy, MCC, F1 Sore, VaR for Rank Error.

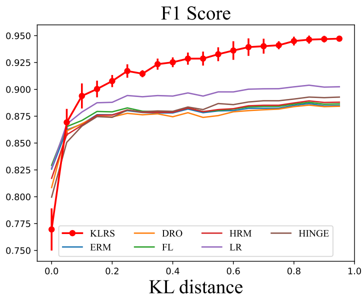
Definition (Rank Error)

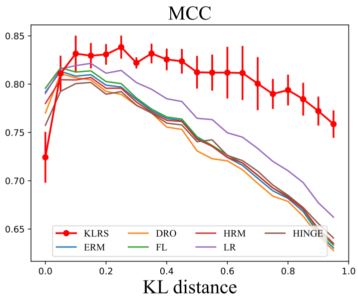
Given a prediction model h, a positive sample x_+ and a negative sample x_-, the ranking error is defined as

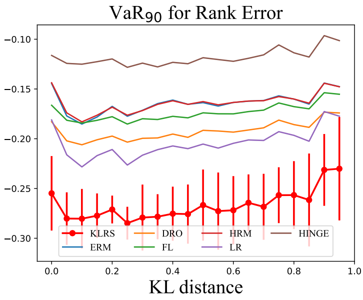
$$\epsilon(h) \triangleq h(x_-) - h(x_+). \tag{11}$$

Experiment Result









The class distribution in the training data is skewed while the test class distribution is evenly distributed.

Problem Settings

- Dataset: generated form CIFAR10 and CIFAR100 by Long-Tail Strategy. LT performs downsampling on each class.
- ρ denotes the ratio between the size of the most common and most rare classes. The larger ρ is, the heavier tailed the class distribution is.
- Existing models that use average in-sample performance as optimization criterion can be easily biased towards dominant classes and perform poorly on minority classes.

Group KL-RS

In this scenario, distribution shift only happens at the class level. Let $w = 0$ in hierarchical KL-RS. Let $\tilde{\boldsymbol{z}} = (\tilde{\mathbf{x}}, \tilde{y})$ and $\tilde{g} = \tilde{y}$, our tailored KL-RS can be the following formulation.

$$\begin{aligned} \min_{\boldsymbol{\theta} \in \Theta, \lambda \geq 0} \quad & \lambda \\ \text{s.t.} \quad & \lambda \log \left(\mathbb{E}_{\hat{\mathbb{P}}_{\tilde{y}}} \exp \left(\mathbb{E}_{\hat{\mathbb{P}}_{\tilde{\mathbf{x}}|\tilde{y}}} l(\boldsymbol{\theta}, (\tilde{\mathbf{x}}, \tilde{y}))/\lambda \right) \right) \leq \tau \end{aligned} \tag{12}$$

Experiment Result

Table 1: Results of Long-Tailed Learning

Dataset	ρ	0.1		0.01	
	Algorithm	average acc	worst acc	average acc	worst acc
CIFAR10 (LT)	ERM	74.93(0.90)	65.77(1.55)	52.01(0.63)	14.72(2.28)
	KL-RS	76.28(0.83)	66.43(0.26)	61.80(0.42)	50.32(0.60)
	CVaRDRO	75.25(1.32)	66.83(0.80)	53.66(0.99)	10.46(4.15)
	Focal	73.55(1.23)	62.34(4.52)	50.83(0.79)	14.35(1.72)
	KL-RS Focal	74.81(0.70)	63.81(2.05)	59.30(1.36)	48.98(0.51)
	Ldam	81.86(0.52)	71.60(1.02)	59.61(1.83)	9.46(4.37)
	KL-RS Ldam	**82.87(0.44)**	**75.06(1.17)**	**70.68(0.13)**	**60.96(1.99)**
CIFAR100 (LT)	ERM	40.28(0.75)	1.98(0.99)	24.84(0.49)	0.00(0.00)
	KL-RS	41.86(0.54)	**15.18(1.14)**	27.82(0.63)	0.00(0.00)
	CVaRDRO	39.55(2.34)	2.75(1.06)	24.74(1.53)	0.00(0.00)
	Focal	40.30(0.49)	2.31(1.14)	25.15(0.56)	0.00(0.00)
	KL-RS Focal	40.67(0.82)	13.53(3.02)	27.69(1.15)	0.00(0.00)
	Ldam	42.85(1.14)	0.00(0.00)	27.10(1.03)	0.00(0.00)
	KL-RS Ldam	**45.65(0.79)**	11.55(2.49)	**31.44(1.21)**	0.00(0.00)

Fair PCA

Fair PCA is a novel method that aims to learn low dimensional representations and obtain uniform performance over all subpopulations.

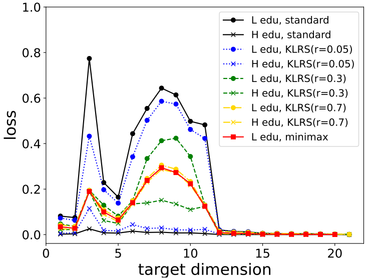
$$\min_{U \in \mathbb{R}^{m \times n}, rank(U) \leq d} \max_{j \in [J]} \left\{ \frac{1}{|A_j|} loss(A_j, A_j U U^T) \right\}.$$

KL-RS PCA aligns with Rawls's *difference principle*. While we allow for different performance levels across subgroups, our optimization goal is to minimize the disparity among all subgroups as much as possible.

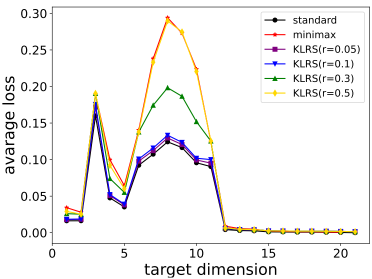
$$\min_{U \in \mathbb{R}^{m \times n}, rank(U) \leq d, \lambda \geq 0} \lambda$$

$$\text{s.t.} \lambda \log \left(\sum_{i=1}^{J} \frac{|A_i|}{m} \exp\left(\frac{1}{\lambda |A_i|} loss(A_i, A_i U U^T)\right) \right) \leq \tau$$

Experiment Result



(a) Performance of different subgroups



(b) Average performance

The End

Questions? Comments?