Distributionally Robust Counterfactual Risk Minimization

Louis Faury, Ugo Tanielian, Elvis Dohmatob, Elena Smirnova, Flavian Vasile
Criteo AI Lab

criteo AI Lab

Key Take Away Message

- We show that DRO provides a **principled** and **general** framework for the CRM problem.
- DRO estimators enjoy asymptotic **consistency** and **performance certificate** guarantees, crucial for CRM.
- We derive a **new** CRM algorithm based on the DRO formulation, **outperforming SOTA** on synthetic datasets.

Offline Policy Optimization

Task

- learning how to act from historical data with **implicit feedback**.
- improve the current version of a search-engine, recommender system (also applications to clinical trials).

Notations

- contexts $x \in \mathcal{X}$ drawn under ν
- actions $y \in \mathcal{A}$ drawn under a policy π
- cost $c(x, y)$ when taking action y to context x

Objective

Minimize the **risk** of the policy π:

$$\min_\pi R(\pi) = \mathbb{E}_{x\sim\nu, y\sim\pi}\left[c(x, y)\right]$$

when the only available data is the **interaction logs** of another policy π_0:

$$\mathcal{H}_0 = \Big(x_i, y_i, p_i = \pi_0(y_i|x_i), c_i = c(x_i, y_i)\Big)_{1\leq i\leq n}$$

To reduce the variance, we prefer the use of clipped propensity scores

$$\min_\pi \hat{R}_n(\pi) = \frac{1}{n}\sum_{i=1}^n c_i \min\left(M, \frac{\pi(x_i|y_i)}{p_i}\right)$$

Challenges and Existing Solutions

Main challenges

- The estimator $\hat{R}_n(\pi)$ can have a very high variance.
- $\hat{R}_n(\pi)$ **does not** provide a performance certificate:

$$\hat{R}_n(\pi) \overset{?}{\leqslant} R(\pi) \text{ w.h.p}$$

$\Rightarrow$ This makes the naive estimator hazardous in practice.

Existing solution

- Counterfactual Risk Minimization (POEM, Swaminathan et al, 2015):

$$\min_\pi \hat{R}_n^\lambda = \hat{R}_n(\pi) + \lambda\sqrt{\widehat{\mathrm{Var}}_n(\pi)/n}$$

  with $\widehat{\mathrm{Var}_n(\pi)}$ is the empirical variance of the counterfactual costs.
- Provides a variance-dependant, consistent performance certificate.
- Can be augmented with **variance-reduction** techniques (Dudik & al 2011, Swaminathan & Joachims, 2015b), also covered by our work.

Distributionally Robust Optimization (DRO)

- Let introduce $\ell(\xi, \theta) = c(x, y)\min\left(M, \frac{\pi_\theta(y|x)}{\pi_0(y|x)}\right)$ and $P = \nu \otimes \pi$.
- DRO treats the empirical distribution $\hat{P}_n$ with **skepticism**:

$$\tilde{R}_n^{\mathcal{U}}(\theta, \epsilon) \triangleq \max_{Q\in\mathcal{U}_\epsilon(\hat{P}_n)} \mathbb{E}_{\xi\sim Q}[\ell(\xi;\theta)].$$

  where $\mathcal{U}_\epsilon(\hat{P}_n)$ is a distributional ambiguity set around $\hat{P}_n$.
- For ambiguity sets based on **coherent φ-divergence**, DRO estimators enjoy nice asymptotic guarantees for CRM (see below).
- POEM is a particular instance of DRO, with χ^2 divergence.

$\Rightarrow$ DRO therefore provides a **general, principled** framework for CRM.

DRO (ctd')

DRO: a general and principled framework for CRM

- **Performance certificate**:

$$\lim_{n\to\infty} \mathbb{P}\left(R(\pi) \leq \tilde{R}_n^{\varphi}(\pi, \epsilon_{n,\delta})\right) \geq 1 - \delta$$

- **Variance penalization**:

$$\tilde{R}_n^{\varphi}(\pi, \epsilon_n) = \hat{R}_n(\pi) + \sqrt{\epsilon_n V_n(\pi)} + o\left(\frac{1}{\sqrt{n}}\right)$$

KL-CRM Algorithms

DRO with **KL-divergence uncertainty sets**:

$$\min_\pi \max_{KL(Q||\hat{P}_n)\leq\varepsilon} \mathbb{E}_{\xi\sim Q}[\ell(\xi;\theta)]$$

- The worst-case distribution takes the form of a Boltzmann distribution
- This leads to minimizing the new CRM objective:

$$\tilde{R}_n^{\text{KL}}(\pi) = \frac{\sum_{i=1}^n \ell(\xi_i;\pi)\exp(\ell(\xi_i;\pi)/\gamma^\star)}{\sum_{j=1}^n \exp(\ell(\xi_j;\pi)/\gamma^\star)}.$$

  ($\gamma^\star$ is an hyperparameter). We call this algorithm **KL-CRM**.
- The optimal temperature γ^* can be approximated:

$$\gamma^* \approx \sqrt{\widehat{\mathrm{Var}}_n(\pi)/2\epsilon}.$$

  γ^* should be updated concurrently to the π during training. We call this algorithm **aKL-CRM**.

Experimental results

We follow the experimental procedure introduced in (Swaminathan et al, 2015). It is a supervised $\to$ unsupervised dataset conversion to build bandit feedback from four multi-label classification datasets. aKL-CRM **equals or outperforms** SOTA.

	Scene	**Yeast**	**RCV1-Topics**	**TMC2009**
CIPS	1.163	4.369	0.929	2.774
POEM	1.157	**4.261**	0.918	2.190
KL-CRM	1.146	4.316	0.922	2.134
aKL-CRM	**1.128**	**4.271**	**0.779**	**2.034**

Table 1: Hamming loss on $\mathcal{D}^*_{\text{test}}$ for the different greedy policies, averaged over 20 independent runs. Bold font indicates that one or several algorithms are statistically better than the rest, according to a one-tailed paired difference t-test at significance level of 0.05.

Another experiment focuses on the impact of the size of the bandit dataset:

- For large datasets, all algorithms confound (as expected).
- For small datasets, the KL-based algorithms outperform POEM.

Future work

- Can we derive **finite sample guarantees** for DRO-based estimators?
- Can other tractable algorithms be derived from the DRO formulation?