Distributionally Robust Counterfactual Risk Minimization

Louis Faury[1,2], Ugo Tanielian[1,3], Elena Smirnova[1]
Flavian Vasile[1], Elvis Dohmatob[1]

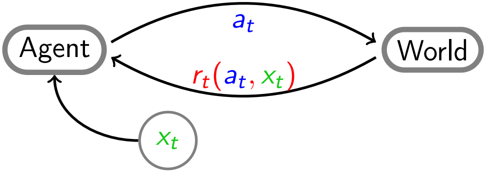
[1] Criteo AI Labs [2] LTCI, Telecom ParisTech [3] LPSM, Paris 6

Outline

- We are interested in **off-line policy evaluation and improvement** in a contextual bandit setting.
- We propose to use tools from **Distributionally Robust Optimization** (DRO) for this task, motivated by asymptotic guarantees.
- We introduce a **new algorithm** for off-line policy improvement, based on the DRO framework, that outperforms the state-of-the-art on classical datasets.

Contextual Bandits (CB)

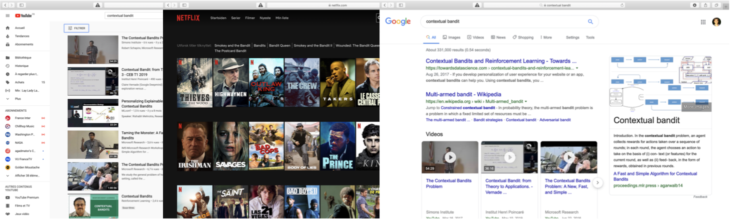
The contextual bandits (CB) is an extension to the classical multi-arm bandit setting.



In CB, an agent is presented with a context x_t (exogenous) and plays an action a_t. The environment then generates a reward r_t.

The agent's goal is to maximize its expected reward.

Contextual Bandit (CB, cntd')



- Recommender system:

$$x_t\text{=user embedding, } a_t\text{=recommandation, } r\text{=click}$$

- Clinical trials:

$$x_t\text{=patient information, } a_t\text{=medication, } r\text{=remission}$$

Contextual Bandit (CB, cntd')

Goal: Maximize the expected reward, under two settings:

- **Online setting**: at every round t, the agents interacts with the world to minimize its cumulative regret. The challenge is the exploration-exploitation trade-off.
- **Offline setting**: the agent only has access to past interactions and must find a way to improve its performance. The challenge is off-line policy evaluation and improvement.

We will consider the offline setting.

Some notations

- Let contexts $x \in \mathcal{X}$ and action $a \in \mathcal{A}$
- Let the cost $c(x, a) := -r(x, a)$.
- The contexts are drawn under ν (unknown).

An agent is characterized by its **policy**: a function that maps contexts to a distribution on the actions.

The goal is to find the policy π with minimal **risk**:

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

which is the expected cost suffered when playing the policy π.

Offline Contextual Bandits (OCB)

In OCB, the agent cannot interact with the environment. The only available data are **interaction logs** from a logging policy π_0:

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

A standard estimator for $R(\pi)$ involves **inverse propensity scores**:

$$R(\pi) = \mathbb{E}_{x \sim \nu, a \sim \pi_0}\left[c(x, a)\frac{\pi(a|x)}{\pi_0(a|x)}\right]$$

usually estimated with capping:

$$\hat{R}_n(\pi) = \frac{1}{n}\sum_{\mathcal{H}_0} c_i \min\left(M, \frac{\pi(a_i|x_i)}{p_i}\right)$$

sometimes called the IPS estimator.

Counterfactual Risk Minimization (CRM)

Problem: the estimator $\hat{R}_n(\pi)$ can have very large variance for some π and may be **over-confident** (optimizer's curse).

Solution: [Swaminathan et al, 2015] [1] suggest looking at an variance-sensitive upper-bound on the true risk:

$$\boxed{R(\pi) \leq \hat{R}_n(\pi) + \lambda \sqrt{\widehat{\mathrm{Var}}_n(\pi)/n} \quad \text{w.h.p}}$$

leading to the CRM principle for policy improvement:

$$\mathrm{argmin}_{\pi} \hat{R}_n(\pi) + \lambda \sqrt{\widehat{\mathrm{Var}}_n(\pi)/n}$$

which gave rise to the **POEM** algorithm (state-of-the-art).

Can be augmented with **variance-reduction** techniques (Self-Normalized estimator, Doubly Robust).

[1] Counterfactual Risk Minimization: Learning from Logged Bandit Feedback

Our contribution

We show that Distributionally Robust Optimization (DRO) tools can be applied to OCB in order to:

- provide a unified framework to build a collection of (asymptotic) **variance-sensitive upper-bounds on the risk**
- derive existing CRM algorithms
- derive new CRM algorithms outperforming state-of-the-art

$\Rightarrow$ DRO provides principled tools for the OCB problem. It is a general framework that generalizes existing CRM solutions.

Distributionally Robust Optimization (DRO)

Denote $\xi := (x, a)$ with distribution $P := \nu \times \pi_0$. We write the empirical risk as follows

$$\hat{R}_n(\pi) = \mathbb{E}_{\xi \sim \hat{P}_n}\left[\ell_\pi(\xi)\right] = \frac{1}{n}\sum_{i=1}^{n} \ell_\pi(\xi_i)$$

where $\ell_\pi(\xi_i) = c_i \min(M, \pi(a_i|x_i)/p_i)$ (capped propensity-costs).

In DRO, we treat $\hat{P}_n$ with **skepticism** and introduce a **robust risk**:

$$\boxed{\tilde{R}_n^{\mathcal{U}}(\pi, \varepsilon) := \sup_{Q \in \mathcal{U}_\varepsilon} \mathbb{E}_{\xi \sim Q}\left[\ell_\pi(\xi)\right]}$$

where $\mathcal{U}_\varepsilon$ is an **ambiguity set**: a «ball» of radius ε around $\hat{P}_n$.

DRO (cnt'd)

We define $\mathcal{U}_\varepsilon$ using coherent φ-divergences

$$\mathcal{U}_\varepsilon = \left\{ Q \ \text{ s.t } D_\varphi(Q||\hat{P}_n) \leq \varepsilon \right\}$$

where for $Q \ll P$:

$$D_\varphi(Q||P) = \int \varphi\left(\frac{dQ}{dP}\right) dP$$

and **1)** φ is a convex function, **2)** $\varphi(t) \geq \varphi(1) = 0$, **3)** $\varphi'(1) = 0$, **4)** $\varphi''(1) > 0$ (coherent conditions).

We will consider robust risk defined through coherent φ-divergences:

$$\boxed{\tilde{R}_n^\varphi(\pi, \varepsilon) = \sup_{D_\varphi(Q||\hat{P}_n) \leq \epsilon} \mathbb{E}_{\xi \sim Q}\left[\ell_\pi(\xi)\right]}$$

DRO for CRM: guarantees

Guarantee 1 The robust risk $\tilde{R}_n^{\varphi}(\pi, \varepsilon)$ provides an **asymptotic performance certificate** for the true risk.

Lemma 1: Risk upper-bound

For any $\delta > 0$:

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

where $\varepsilon_n = \varphi''(1)\chi^2_{1,1-\delta}/(2n)$.

This result can be derived from Proposition 1 in [Duchi 2016][2].

[2] Statistics of Robust Optimization: A Generalized Empirical Likelihood Approach

DRO for CRM: guarantees (2)

Guarantee 2 The robust risk **penalizes high-variance** estimates:

Lemma 2: Asymptotic variance decomposition

$$\tilde{R}_n^{\varphi}(\pi, \varepsilon/n) = \hat{R}_n(\pi) + \sqrt{\frac{\varepsilon}{n}\widehat{\mathrm{Var}}_n(\pi)} + o(\frac{1}{\sqrt{n}})$$

This result can be obtained as a Corollary of Theorem 2 of [Duchi 2016].

$\Rightarrow$ Lemma 1 and Lemma 2 imply that the upper-bounds provide variance-sensitive performance certificate, making it a reliable tool for **off-line policy evaluation**.

DRO for CRM: guarantees (3)

Guarantee 3 With exponentially parametrized policy, minimizing the robust risk with χ^2 ambiguity sets is **exactly** the POEM algorithm.

Lemma 3: Exact variance decomposition

For ε small enough:

$$\tilde{R}_n^{\chi^2}(\pi, \varepsilon) = \hat{R}_n(\pi) + \sqrt{\varepsilon \widehat{\mathrm{Var}}_n(\pi)}$$

$\Rightarrow$ Existing CRM algorithms are already instances of DRO estimators!

DRO for CRM: guarantees (4)

Sketch of proof By strong duality we have

$$\sup_{D_\varphi(Q\|\hat{P}_n)} \mathbb{E}_Q\left[\ell_\pi(\xi)\right] = \inf_{\gamma \geq 0} \gamma\varepsilon + \inf_Q \left\{\mathbb{E}_Q\left[\ell_\pi(\xi)\right] - \gamma D_\varphi(Q\|\hat{P}_n)\right\} \tag{1}$$

Using the Envelope Theorem of [Rockafellar18][3] one gets:

$$\sup_{D_\varphi(Q\|\hat{P}_n)} \mathbb{E}_Q\left[\ell_\pi(\xi)\right] = \inf_{\gamma \geq 0} \gamma\varepsilon + \inf_c \left\{c + \gamma \mathbb{E}_{\hat{P}_n}\left[\varphi^*\left((\ell_\pi(\xi) - c)/\gamma\right)\right]\right\} \tag{2}$$

For the χ^2-divergence, $\varphi(z) = (z-1)^2$ and $\varphi^*(s) = s^2/4 + s$ for $s \geq -2$. Solving leads to the result.

[3] Risk and utility in the duality framework of convex analysis

DRO for CRM: guarantees (5)

Sum-up so far

- Guarantees 1 and 2: DRO is a general tool for building variance-sensitive upper-bounds on the risk
- Guarantee 3: POEM is actually DRO with χ^2 divergences.

In what follows

We introduce a **new** CRM algorithm inspired from DRO, and derived from Kullback-Leibler divergence ambiguity sets.

New KL-based CRM algorithm

We now consider **KL-based** ambiguity sets:

$$\tilde{R}_n^{\mathrm{KL}}(\pi, \varepsilon) = \min_{KL(Q||\hat{P}_n)} \mathbb{E}_{\xi \sim Q}\left[\ell_\pi(\xi)\right]$$

There is a **tractable computation** for the worst-case distribution.

Lemma 4: KL robustified risk

It exists $\gamma > 0$ such that

$$\tilde{R}_n^{\mathrm{KL}}(\pi, \varepsilon) = \sum_{i=1}^{n} \frac{\ell_\pi(\xi) e^{\ell_\pi(\xi_i)/\gamma}}{\sum_{j=1}^{n} e^{\ell_\pi(\xi_j)/\gamma}} \tag{3}$$

The line of proof follows the one of Lemma 3, and uses the convex conjugate of $\varphi_{\mathrm{KL}}(z) = z\log(z) - z + 1$.

New CRM algorithms

Policy optimization: Minimize the upper-bound given by the robust risk! This gives rise to the KL-CRM algorithm:

$$\text{minimize}_{\pi} \left[\tilde{R}_n^{\mathrm{KL}}(\pi, \varepsilon) = \sum_{i=1}^{n} \frac{\ell_\pi(\xi_i) e^{\ell_\pi(\xi_i)/\gamma}}{\sum_{j=1}^{n} e^{\ell_\pi(\xi_j)/\gamma}} \right] \qquad \text{(KL-CRM)}$$

where γ is treated as a hyper-parameter (cross-validation).

Temperature γ dictates the level of pessimism:

- $\gamma \to \infty$ reduces to the IPS estimator
- $\gamma \to 0$ only consider the worst case propensity cost.

New CRM algorithms

A finer analysis reveals a good approximation γ.

Lemma 5: aKL-CRM

$$\gamma_* \simeq \sqrt{\frac{\widehat{Var}_n(\pi)}{2\varepsilon}}$$

The proof relies on a second-order Taylor approximation of the log-m.g.f of the loss in the dual objective.

This gives rise to **aKL-CRM** which minimizes the KL-CRM objective and concurrently updates γ_*.

Experimental Results

We evaluate on **standard datasets** (supervised$\rightarrow$bandit) and compare KL-CRM and aKL-CRM with the basic IPS approach and the POEM algorithm.

Hyper-parameters are determined through cross-validation. Experiments are average over 20 different random initialization.

The performance of a policy is reported by its **expected** instant regret or by the instant regret of its **greedy** policy.

Experimental Results (ctn'd)

Expected instant regret:

	Scene	Yeast	RCV1-Topics	TMC2009
π_0	1.529	5.542	1.462	3.435
CIPS	1.163	4.658	0.930	2.776
POEM	1.157	**4.535**	0.918	2.191
KL-CRM	1.146	4.604	0.922	2.136
aKL-CRM	**1.128**	**4.553**	**0.783**	**2.126**
CRF	0.646	2.817	0.341	1.187

Table: Expected Hamming loss on $\mathcal{D}^*_{\text{test}}$ for the different algorithms, averaged over 20 independent runs. Bold font indicate 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.

Rq: CRF is a **skyline** that has access to full supervised feedback.

Experimental Results (ctn'd)

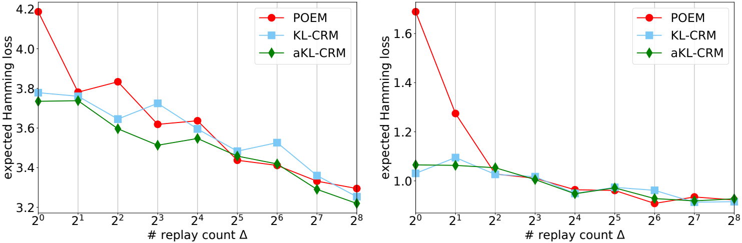
Greedy instant regret:

	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: 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.

Experimental Results (ctn'd)

Influence of the size of the logged history:



(a) Yeast dataset (b) Scene dataset

Figure: Impact of the replay count Δ on the expected Hamming loss. Results are average over 10 independent runs, that is 10 independent train/test split and bandit dataset creation. KL-CRM and aKL-CRM outperform POEM in the small data regime.

Conclusion and future work

DRO is a principled tool for OCB and lead to competitive CRM algorithms.

Future work:

- further experimental evalutions (SNIPS, DR)
- solving the primal problem can be easy! we can use performance certificate given by many φ divergences.
- can we derive finite samples guarantees?

Thank you!