Distributionally Robust Counterfactual Risk Minimization

Louis Faury,[1,2,*] Ugo Tanielian,[1,3,*] Flavian Vasile,[1] Elena Smirnova,[1] Elvis Dohmatob[1]
[1]Criteo AI Lab
[2]LTCI, Telecom-ParisTech, Université Paris-Saclay
[3] LPSM, Université Paris 6
{l.faury, u.tanielian, f.vasile, e.smirnova, e.dohmatob}@criteo.com

Abstract

This manuscript introduces the idea of using Distributionally Robust Optimization (DRO) for the Counterfactual Risk Minimization (CRM) problem. Tapping into a rich existing literature, we show that DRO is a principled tool for counterfactual decision making. We also show that well-established solutions to the CRM problem like sample variance penalization schemes are special instances of a more general DRO problem. In this unifying framework, a variety of distributionally robust counterfactual risk estimators can be constructed using various probability distances and divergences as uncertainty measures. We propose the use of Kullback-Leibler divergence as an alternative way to model uncertainty in CRM and derive a new robust counterfactual objective. In our experiments, we show that this approach outperforms the state-of-the-art on four benchmark datasets, validating the relevance of using other uncertainty measures in practical applications.

1 Introduction

Learning how to act from historical data is a largely studied field in machine learning (Strehl et al. 2010; Dudík, Langford, and Li 2011; Li et al. 2011; 2015), spanning a wide range of applications where a system interacts with its environment (e.g search engines, ad-placement and recommender systems). Interactions are materialized by the actions taken by the system, themselves rewarded by a feedback measuring their relevance. Both quantities can be logged at little cost, and subsequently used to improve the performance of the learning system. The Batch Learning from Bandit Feedback (Swaminathan and Joachims 2015a; 2015b) (BLBF) framework describes such a situation, where a *contextual* decision making process must be improved based on the logged history of *implicit* feedback observed only on a subset of actions. Counterfactual estimators (Bottou et al. 2013) allow to forecast the performance of any system from the logs, as if it was taking the actions by itself. This enables the search for an optimal system, even with observations biased towards actions favored by the logger.

A natural approach to carry out this search consists in favoring systems that select actions with high empirical counterfactual rewards. However, this initiative can be rather burdensome as it suffers a crucial caveat intimately linked with a phenomenon known as the optimizer's curse (Capen et al. 1971; Smith and Winkler 2006; Thaler 2012). It indicates that sorting actions by their empirical reward average can be sub-optimal since the resulting expected *post-decision surprise* is non-zero. In real life applications where the space of possible actions is often extremely large and where decisions are taken based on very low sample sizes, the consequence of this phenomenon can be dire and motivates the design of principled *robust* solutions. As a solution for this Counterfactual Risk Minimization (CRM) problem, the authors of (Swaminathan and Joachims 2015a) proposed a modified action selection process, penalizing behaviors resulting in high-variance estimates.

In this paper, we argue that another line of reasoning resides in using Distributional Robust Optimization (DRO) for the CRM. It has indeed proven to be a powerful tool both in decision theory (Duchi, Glynn, and Namkoong 2016; Esfahani and Kuhn 2018; Blanchet and Murthy 2019) and the training of robust classifiers (Madry et al. 2017; Xiao et al. 2018; Hu et al. 2018). Under the DRO formulation, one treats the empirical distribution with skepticism and hence seeks a solution that minimizes the worst-case expected cost over a family of distributions, described in terms of an uncertainty ball. Using distributionally robust optimization, one can therefore control the magnitude of the post-decision surprise, critical to the counterfactual analysis.

We motivate the use of DRO for the CRM problem with asymptotic guarantees and bring to light a formal link between the variance penalization solution of (Swaminathan and Joachims 2015a) and a larger DRO problem for which the uncertainty set is defined with the chi-square divergence. Building from this, we propose the use of other uncertainty sets and introduce a KL-based formulation of the CRM problem. We develop a new algorithm for this objective and benchmark its performance on a variety of real-world datasets. We analyze its behavior and show that it outperforms existing state-of-the-art methods.

The structure of the paper is the following: in Section 2 we formally introduce the BLBF framework and the CRM problem. In Section 3 we present the DRO framework, moti-

[*]Equal contribution.

vate it for CRM and re-derive the POEM (Swaminathan and Joachims 2015a) algorithm as one of its special cases and introduce a new CRM algorithm. In Section 4 we compare this new algorithm with state-of-the-art CRM algorithms on four public datasets and finally summarize our findings in Section 5.

2 Batch Learning from Logged Bandit Feedback

2.1 Notation and terminology

We denote $x \in \mathcal{X}$ arbitrary *contexts* drawn from an unknown distribution ν and presented to the decision maker. Such a quantity can describe covariate information about a patient for a clinical test, or a potential targeted user in a recommender system. The variable $y \in \mathcal{Y}$ denotes the *actions* available to the decision maker - the potential medications to give to the patient, or possible advertisements targeting the user for instance. A *policy* is a mapping $\pi : \mathcal{X} \to \mathcal{P}(\mathcal{Y})$ from the space of contexts to probabilities in the action space. For a given (context, action) pair (x, y), the quantity $\pi(y|x)$ denotes the probability of the policy π to take the action y when presented the context x. When picking the action y for a given context x, the decision maker receives a *reward* $\delta(x, y)$, drawn from an unknown distribution. In our examples, this reward could indicate a patient's remission, or the fact that the targeted user clicked on the displayed ad. This reward $\delta(x, y)$ can also be assumed to be deterministic - as in (Swaminathan and Joachims 2015a; 2015b). We make this assumption in the rest of this manuscript. Finally, for a given context $x \in \mathcal{X}$ and an action $y \in \mathcal{Y}$, we define the cost function $c(x, y) \triangleq -\delta(x, y)$.

In this paper, we try to find policies producing low expected costs. To make this search tractable, it is usual to restrict the search to a family of *parametric* policies, henceforth tying policies π_θ to a vector $\theta \in \Theta$. The *risk*:

$$R(\theta) \triangleq \mathbb{E}_{x\sim\nu, y\sim\pi_\theta(\cdot|x)}\left[c(x, y)\right] \tag{1}$$

corresponds to the expected cost obtained by the policy π_θ through the different actions y taken, a quantity the decision maker will try to minimize.

2.2 Counterfactual Risk Minimization

In practical applications, it is common that one has only access to the *interaction logs* of a previous version of the decision making system, also called a *logging policy* (denoted π_0). More formally, we are interested in the case where the only available data is a collection of quadruplets $\mathcal{H} \triangleq (x_i, y_i, p_i, c_i)_{1\leq i\leq n}$, where the costs $c_i \triangleq c(x_i, y_i)$ were obtained after taking an action y_i with probability $p_i \triangleq \pi_0(y_i|x_i)$ when presented with a context $x_i \sim \nu$.

In order to search for policies π_θ with smaller risk than π_0, one needs to build counterfactual estimators for $R(\theta)$ from the historic $\mathcal{H}$. One way to do so is to use *inverse propensity scores* (Rosenblum and Rubin 1983):

$$\begin{aligned} R(\theta) &= \mathbb{E}_{x\sim\nu, y\sim\pi_\theta(x)}\left[c(x, y)\right] \\ &= \mathbb{E}_{x\sim\nu, y\sim\pi_0(x)}\left[c(x, y)\frac{\pi_\theta(y|x)}{\pi_0(y|x)}\right] \end{aligned} \tag{2}$$

for any π_θ absolutely continuous w.r.t π_0. Henceforth, $R(\theta)$ can be approximated with samples (x_i, y_i, p_i, c_i) from the interaction logs $\mathcal{H}$ via the sample average approximation:

$$R(\theta) \simeq \frac{1}{n}\sum_{i=1}^{n} c_i \frac{\pi_\theta(y_i|x_i)}{p_i} \tag{3}$$

Bluntly minimizing the objective provided by the counterfactual risk estimator (3) is known to be sub-optimal, as it can have *unbounded* variance (Ionides 2008). It is therefore a classical technique (Bottou et al. 2013; Cortes, Mansour, and Mohri 2010; Strehl et al. 2010; Swaminathan and Joachims 2015a) to *clip* the propensity weights . This leads to the Clipped Inverse Propensity Scores (CIPS) estimator:

$$\hat{R}_n(\theta) \triangleq \frac{1}{n}\sum_{i=1}^{n} c_i \min\left(M, \frac{\pi_\theta(y_i|x_i)}{p_i}\right). \tag{4}$$

The variable M is an hyper-parameter, balancing the variance reduction brought by weight clipping and the bias introduced in the empirical estimation of $R(\theta)$. The search for a minimal θ with respect to the estimator $\hat{R}_n(\theta)$ is often referred to as *Counterfactual Risk Minimization* (CRM).

Remark 1 Other techniques have been proposed in the literature to reduce the variance of the risk estimators. For instance, the doubly robust risk (Dudík, Langford, and Li 2011) takes advantage of both counterfactual estimators and supervised learning methods, while the self-normalized risk (Swaminathan and Joachims 2015b) was designed to counter the effect of a phenomenon known as propensity overfitting. We do not explicitly cover them in our analysis, but the results we derive hereinafter also hold for such estimators.

2.3 Sample-Variance Penalization

The main drawback of the CIPS estimator is that two different policies can have risk estimates of highly different variance - something the sample average approximation cannot capture. The authors of (Swaminathan and Joachims 2015a) developed a variance-sensitive action selection process where one penalizes policies with high-variance risk estimates. Their approach is based on a sample-variance penalized version of the CIPS estimator:

$$\hat{R}_n^\lambda(\theta) \triangleq \hat{R}_n(\theta) + \lambda\sqrt{V_n(\theta)/n}, \tag{5}$$

where λ is an hyper-parameter set by the practitioner, and $V_n(\theta)$ denotes the empirical variance of the quantities $c_i \min\left(M, \frac{\pi_\theta(y_i|x_i)}{p_i}\right)$. The main motivation behind this approach is based on confidence bounds derived in (Maurer and Pontil 2009), upper-bounding with high-probability the true risk $R(\theta)$ by the empirical risk $\hat{R}_n(\theta)$ augmented with an additive empirical variance term. In a few words, this allows to build and optimize a pessimistic envelope for the true risk and penalize policies with high variance risk estimates. The authors of (Swaminathan and Joachims 2015a) proposed the Policy Optimization for Exponential Models (POEM) algorithm and showed state-of-the-art results on a collection of counterfactual tasks when applying this method to exponentially parametrized policies:

$$\pi_\theta(y|x) \propto \exp\left(\theta^T \phi(x,y)\right), \tag{6}$$

with $\phi(x,y)$ being a d-dimensional joint feature map.

3 Distributionally Robust Counterfactual Risk Minimization

3.1 Motivating distributional robustness for CRM

For more concise notations, let us introduce the variable $\xi = (x,y)$, the distribution $P = \nu \otimes \pi_0$ and the loss $\ell(\xi,\theta) \triangleq c(x,y) \min\left(M, \frac{\pi_\theta(y|x)}{\pi_0(y|x)}\right)$. The minimization of the counterfactual risk (2), now writes $\theta^\star \triangleq \operatorname{argmin}_{\theta\in\Theta} \mathbb{E}_{\xi\sim P}\left[\ell(\xi,\theta)\right]$ and its empirical counterpart $\hat{\theta}_n \triangleq \operatorname{argmin}_{\theta\in\Theta} \hat{R}_n(\theta)$. The consistency of the estimator $\hat{\theta}_n$ holds under general conditions and the empirical risk converges to the optimal true risk (Vapnik 1992):

$$\mathbb{E}_{\xi\sim P}[\ell(\xi;\theta^\star)] - \hat{R}_n(\hat{\theta}_n) \underset{n\to\infty}{\longrightarrow} 0 \tag{7}$$

However, one of the major drawbacks of this approach is that the empirical risk $\hat{R}_n(\theta)$ cannot be used as a *performance certificate* for the true risk $\mathbb{E}_{\xi\sim P}[\ell(\xi;\theta)]$. Indeed, one will fail at controlling the true risk of any parameter θ since by the Central Limit Theorem:

$$\lim_{n\to\infty} \mathbb{P}\left(\mathbb{E}_{\xi\sim P}[\ell(\xi;\theta)] \leq \hat{R}_n(\theta)\right) = 1/2 \tag{8}$$

One way to circumvent this limitation is to treat the empirical distribution $\hat{P}_n$ with *skepticism* and to replace it with an uncertainty set $\mathcal{U}_\epsilon(\hat{P}_n)$ of distributions around $\hat{P}_n$ with $\epsilon > 0$ a parameter controlling the size of the uncertainty set $\mathcal{U}_\epsilon(\hat{P}_n)$. This gives rise to the *distributionally robust* counterfactual risk:

$$\tilde{R}_n^{\mathcal{U}}(\theta,\epsilon) \triangleq \max_{Q\in\mathcal{U}_\epsilon(\hat{P}_n)} \mathbb{E}_{\xi\sim Q}[\ell(\xi;\theta)]. \tag{9}$$

Minimizing this quantity w.r.t to θ yields the general DRO program:

$$\begin{aligned}\tilde{\theta}_n &\triangleq \operatorname{argmin}_{\theta\in\Theta} \tilde{R}_n^{\mathcal{U}}(\theta,\epsilon) \\ &= \underset{\theta\in\Theta}{\operatorname{argmin}} \max_{Q\in\mathcal{U}_\epsilon(\hat{P}_n)} \mathbb{E}_{\xi\sim Q}[\ell(\xi;\theta)].\end{aligned} \tag{10}$$

There is liberty on the way to construct the uncertainty set $\mathcal{U}_\epsilon(\hat{P}_n)$ including parametric (Madry et al. 2017; Xiao et al. 2018) and non-parametric designs (Parys, Esfahani, and Kuhn 2017; Sinha, Namkoong, and Duchi 2017; Blanchet and Murthy 2019). Moreover, for well chosen uncertainty sets (Duchi, Glynn, and Namkoong 2016), one can prove performance guarantees asserting that asymptotically (in the limit $n\to\infty$), for all $\theta\in\Theta$:

$$\begin{aligned} &\mathbb{E}_{\xi\sim P}[\ell(\xi;\theta)] \leq \tilde{R}_n^{\mathcal{U}}(\theta,\epsilon_n) \text{ w.h.p} \\ \text{and}\quad &\mathbb{E}_{\xi\sim P}[\ell(\xi;\theta)] - \tilde{R}_n^{\mathcal{U}}(\theta,\epsilon_n) \to 0 \end{aligned}$$

The robust risk therefore acts as a consistent certificate on the true risk. We believe that these properties alone are enough to motivate the use of DRO for the CRM problem, as it provides an elegant way to *design consistent asymptotic upper bounds for the true risk* and ensure a small post-decision surprise. We detail such guarantees in the next subsection. Later, we draw links between DRO and the POEM, showing that a wide variety of DRO problems *account for the empirical variance* of the samples, therefore mitigating the limitations of empirical averages as discussed in Section 2.2.

3.2 Guarantees of robustified estimators with φ-divergences

We are interested in DRO instances that are amenable to direct optimization. To this end, we focus here only on uncertainty sets $\mathcal{U}_\epsilon(\hat{P}_n)$ based on information divergences (Csiszár 1967), since they strike a nice compromise between ease of implementation and theoretical guarantees that will reveal useful for the CRM problem. The use of information divergences for DRO has already been largely studied in several works (for example (Duchi, Glynn, and Namkoong 2016; Gotoh, Kim, and Lim 2018)). For the sake of completeness, we now recall in Definition 1 the definition of information divergences.

Definition 1 (φ-divergences). *Let φ be a real-valued, convex function such that $\varphi(1) = 0$. For a reference distribution P, the divergence of another distribution Q with respect to P is defined by*

$$D_\varphi(Q\|P) \triangleq \begin{cases} \int \varphi(dQ/dP)dP, & \textit{if } Q \ll P, \\ +\infty, & \textit{else.} \end{cases} \tag{11}$$

Subsequently, the definition of the uncertainty set $\mathcal{U}_\epsilon$ relies only on φ-divergences as follows:

$$\mathcal{U}_\epsilon(\hat{P}_n) = \left\{Q \mid D_\varphi(Q\|\hat{P}_n) \leq \epsilon\right\}. \tag{12}$$

We need to ensure that the set of φ-divergences used to define the resulting robust risk $\tilde{R}_n^\varphi(\theta,\epsilon)$ satisfies some basic *coherence* properties. We therefore make further assumptions about the measure of risk φ to narrow the space of information divergences we consider:

Assumption 1 (Coherence). *φ is a real-valued function satisfying:*

- *φ is convex and lower-semi-continuous*
- *$\varphi(t) = \infty$ for $t < 0$, and $\varphi(t) \geq \varphi(1) = 0, \ \forall t \in \mathbb{R}$*
- *φ is twice continuously differentiable at $t = 1$ with $\varphi'(1) = 0$ and $\varphi''(1) > 0$*

The axioms presented in Assumption 1 have been proposed and studied extensively in (Rockafellar 2018). Examples of coherent divergences include the Chi-Square, Kullback-Leibler divergences and the squared Hellinger distance.

Before stating the announced asymptotic guarantees, we make further assumptions on the structure of both the context and parameter spaces.

Assumption 2 (Structure).

- *Θ is a compact subset of some $\mathbb{R}^d$.*

- *$\mathcal{X}$ is a compact subset of some $\mathbb{R}^D$.*

We now state Lemma 1 which provides asymptotic certificate guarantees for the robust risk, asymptotically controlling the true counterfactual risk $\mathbb{E}_{\xi\sim P}[\ell(\xi;\theta)]$ with high probability. It is easy to show that under Assumptions 1 and 2, Lemma 1 can be obtained by a direct application of Proposition 1 of (Duchi, Glynn, and Namkoong 2016).

Lemma 1 (Asymptotic guarantee - Proposition 1 of (Duchi, Glynn, and Namkoong 2016))**.** *Under Assumptions 1 and 2, for a fixed level of confidence $\delta \in (0,1]$, we have that $\forall\theta\in\Theta$:*

$$\lim_{n\to\infty} \mathbb{P}\left(\mathbb{E}_{\xi\sim P}\left[\ell(\xi;\theta)\right] \leq \tilde{R}_n^{\varphi}(\theta,\epsilon_n)\right) \geq 1-\delta \qquad (13)$$

where ϵ_n is defined by the $(1-\delta)$ Chi-Squared quantile, $\epsilon_n(\delta) = \chi^2_{1,1-\delta}/n$.

Proof. The proof mainly consists in showing that under Assumptions 1 and 2, the conditions for applying the Proposition 1 of (Duchi, Glynn, and Namkoong 2016) are fulfilled. □

For the CRM problem, this result is of upmost importance as it allows us to control the post-decision surprise suffered for a given policy π_θ and allows for a pointwise control of the true risk.

Remark 2 A stronger result than Lemma 1 would control with high probability the true risk of the robustified policy $\mathbb{E}_{\xi\sim P}[\ell(\xi;\tilde{\theta}_n)]$ with the optimal value $\tilde{R}_n^{\mathcal{U}}(\tilde{\theta}_n,\epsilon_n)$. It is easy to see that, if $P\in\mathcal{U}_\epsilon$, then $\mathbb{E}_{\xi\sim P}[\ell(\xi;\tilde{\theta}_n)] \leq \tilde{R}_n^{\mathcal{U}}(\tilde{\theta}_n,\epsilon_n)$, hence $\mathbb{P}(\mathbb{E}_{\xi\sim P}[\ell(\xi;\hat{\theta}_n)] \leq \tilde{R}_n^{\mathcal{U}}(\tilde{\theta}_n,\epsilon_n)) \geq \mathbb{P}(P\in\mathcal{U}_\epsilon)$. By exhibiting strong rates of convergence of the empirical distribution $\hat{P}_n$ towards the true distribution P, such a result could be reached. Under mild assumptions on P, this guarantee has been proved in (Esfahani et al. 2017) for Wasserstein based uncertainty sets. In our current case where $\mathcal{U}_\epsilon$ is defined by information divergences this result holds solely under the assumption that P is finitely supported (Van Parys, Esfahani, and Kuhn 2017), a plausible situation when the logging policy is defined on a finite number of (context, action) pairs.

3.3 Equivalences between DRO and SVP

In this subsection, we focus on stressing the link between DRO and sample variance penalization schemes as used in the POEM algorithm. In Lemma 2, we present an asymptotic equivalence between the robust risk (defined with coherent φ divergences) and the SVP regularization used in POEM. This Lemma is a specific case of existing results, already detailed in (Duchi, Glynn, and Namkoong 2016; Namkoong and Duchi 2017) and (Gotoh, Kim, and Lim 2017; 2018).

Lemma 2. *[Asymptotic equivalence - Theorem 2 of (Duchi, Glynn, and Namkoong 2016)] Under Assumptions 1 and 2, for any $\epsilon \geq 0$, integer $n>0$ and $\theta\in\Theta$ we have:*

$$\tilde{R}_n^{\varphi}(\theta,\epsilon_n) = \hat{R}_n(\theta) + \sqrt{\epsilon_n V_n(\theta)} + \alpha_n(\theta), \qquad (14)$$

with $\sup_\theta \sqrt{n}|\alpha_n(\theta)| \xrightarrow{P} 0$ and $\epsilon_n = \epsilon/n$.

Proof. The proof is rather simple, as one only needs to show that the assumptions behind Theorem 2 of (Duchi, Glynn, and Namkoong 2016) are satisfied. □

This expansion gives intuition on the practical effect of the DRO approach: namely, it states that the minimization of the robust risk $\tilde{R}_n^{\varphi}(\theta)$ based on coherent information divergences is *asymptotically equivalent* to the POEM algorithm. This link between POEM and DRO goes further: the following Lemma states that sample-variance penalization is an *exact* instance of the DRO problem when the uncertainty set is based on the chi-square divergence.

Lemma 3 (Non-asymptotic equivalence)**.** *Under Assumption 2, for χ^2-based uncertainty sets, for any $\epsilon\geq 0$, integer $n>0$ and $\theta\in\Theta$ we have:*

$$\tilde{R}_n^{\chi^2}(\theta,\epsilon) = \hat{R}_n(\theta) + \sqrt{\epsilon V_n(\theta)}. \qquad (15)$$

Proof. The line of proof is similar to the proof of Theorem 3.2 of (Gotoh, Kim, and Lim 2018). To ease notations, we denote $Z \triangleq \ell(\xi,\theta)$. Let's consider φ a coherent information divergence and $\varphi^*(z) \triangleq \sup_{t>0} zt - \varphi(t)$ its *convex conjugate*. By strong duality:

$$\sup_{D_\varphi(Q\|\hat{P}_n)\leq\epsilon} \mathbb{E}_Q[Z] = \inf_{\gamma\geq 0} \gamma\epsilon + \sup_Q \left(\mathbb{E}_Q[Z] - \gamma D_\varphi(Q\|\hat{P}_n)\right)$$

Theorem 4 of (Rockafellar 2018) states that:

$$\sup_Q \left(\mathbb{E}_Q[Z] - \gamma D_\varphi(Q\|\hat{P}_n)\right) = \mathbb{E}_{\hat{P}_n}[Z] + \inf_{c\in\mathbb{R}} (c + \gamma\mathbb{E}_{\hat{P}_n}[\varphi^*((Z-c)/\gamma)])$$

Hence we obtain that:

$$\sup_{D_\varphi(Q\|\hat{P}_n)\leq\epsilon} \mathbb{E}_Q[Z] = \inf_{\gamma\geq 0} \gamma\epsilon + \mathbb{E}_{\hat{P}_n}[Z] + \inf_{c\in\mathbb{R}} (c + \gamma\mathbb{E}_{\hat{P}_n}[\varphi^*((Z-c)/\gamma)])$$

In the specific case of the modified χ^2 divergence, $\varphi(z) = \frac{1}{2\sqrt{2}}(z-1)^2$ and its convex conjugate is $\varphi^\star(z) = z + z^2$. Solving $\inf_{c\in\mathbb{R}}(c + \gamma\mathbb{E}_{\hat{P}_n}[\varphi^*((Z-c)/\gamma)])$ leads to:

$$\sup_{D_\varphi(Q\|\hat{P}_n)\leq\epsilon} \mathbb{E}_Q[Z] = \mathbb{E}_{\hat{P}_n}[Z] + \inf_{\gamma\geq 0}\left(\gamma\epsilon + \frac{1}{\gamma}V_n(Z)\right) = \mathbb{E}_{\hat{P}_n}[Z] + \sqrt{\epsilon V_n(Z)}$$

hence the announced result. □

In the case of χ^2-based uncertainty sets, the expansion of Lemma 2 holds non-asymptotically, and the POEM algorithm can be interpreted as the minimization of the distributionally robust risk $\tilde{R}_n^{\chi^2}(\theta,\epsilon)$. To the best of our knowledge, it is the first time that a finite sample equivalence is established between counterfactual risk minimization with SVP (Swaminathan and Joachims 2015a) and DRO.

3.4 Kullback-Leibler based CRM

Among information divergences, we are interested in the ones that allow tractable optimization. Going towards this direction, we investigate in this subsection the robust counterfactual risk generated by Kullback-Leibler (KL) uncertainty sets and stress its efficiency. The KL divergence is a *coherent* φ-divergence, with $\varphi(z) \triangleq z\log(z) + z - 1$ for $z > 0$ and $\varphi(z) = \infty$ elsewhere. The robust risk $\tilde{R}_n^{\text{KL}}(\theta, \epsilon)$ therefore benefits from the guarantees of Lemma 1. Furthermore, it enjoys a simple analytic formula stated in Lemma 4 that allows for direct optimization.

Lemma 4 (Kullback-Leibler Robustified Counterfactual Risk)**.** *Under Assumption 2, the robust risk defined with a Kullback-Leibler uncertainty set can be rewritten as:*

$$\tilde{R}_n^{\text{KL}}(\theta, \epsilon) = \inf_{\gamma>0}\Big(\gamma\epsilon + \gamma\log\mathbb{E}_{\xi\sim\hat{P}_n}[\exp(\ell(\xi;\theta)/\gamma)]\Big) \tag{16}$$

$$= \mathbb{E}_{\xi\sim\hat{P}_n^{\gamma^\star}(\theta)}[\ell(\xi;\theta)]. \tag{17}$$

where $\hat{P}_n^\gamma$ denotes the Boltzmann distribution at temperature $\gamma > 0$, defined by

$$\hat{P}_n^\gamma(\xi_i|\theta) = \frac{\exp(\ell(\xi_i;\theta)/\gamma)}{\sum_{j=1}^n \exp(\ell(\xi_j;\theta)/\gamma)}.$$

Proof. To ease notations, we denote $Z \triangleq \ell(\xi, \theta)$. As in the proof of Lemma 3, one can obtain that:

$$\max_{KL(Q||\hat{P}_n)\leq\epsilon} \mathbb{E}_Q[\ell(\xi;\theta)] = \inf_{\gamma\geq 0}\Big\{\gamma\epsilon + \inf_{c\in\mathbb{R}}\Big(c + \gamma\mathbb{E}_{\hat{P}_n}[\varphi^*_{\text{KL}}((Z-c)/\gamma)]\Big)\Big\}$$

Since $\varphi_{\text{KL}}(z) = z\log z - z + 1$, it is a classical convex analysis exercise to show that its convex conjugate is $\varphi^*_{\text{KL}}(z) = e^z - 1$. Therefore:

$$\max_{KL(Q||\hat{P}_n)\leq\epsilon} \mathbb{E}_Q[\ell(\xi;\theta)] = \inf_{\gamma\geq 0}\Big\{\gamma\epsilon + \inf_{c\in\mathbb{R}}\Big(c + \gamma\mathbb{E}_{\hat{P}_n}[e^{(Z-c)/\gamma} - 1]\Big)\Big\}$$

Solving

$$\inf_{c\in\mathbb{R}}\Big(c + \gamma\mathbb{E}_{\hat{P}_n}[e^{(Z-c)/\gamma} - 1]\Big)$$

is straightforward and leads to:

$$\max_{KL(Q||\hat{P}_n)\leq\epsilon} \mathbb{E}_Q[\ell(\xi;\theta)] = \inf_{\gamma\geq 0}\Big\{\gamma\epsilon + \gamma\log\mathbb{E}_{\hat{P}_n}\Big[e^{Z/\gamma}\Big]\Big\} \tag{18}$$

Differentiating the r.h.s and setting to 0 shows that the optimal γ is solution to the fixed-point equation:

$$\gamma = \frac{\mathbb{E}_{\hat{P}_n^\gamma}[Z]}{\epsilon + \log(\mathbb{E}_{\hat{P}_n}[e^{Z/\gamma}])} \tag{19}$$

where $d\hat{P}_n^\gamma(z) := (1/\mathbb{E}_{\hat{P}_n}[e^{z/\gamma}])e^{z/\gamma}d\hat{P}_n(z)$ is the density of the Gibbs distribution at temperature γ and state degeneracies $\hat{P}_n$. Replacing this value for γ in the r.h.s of (18) yields the announced result. This formula has also been obtained in (Hu and Hong 2013), using another line of proof. □

A direct consequence of Lemma 4 is that the worst-case distribution in the uncertainty set (defined by the KL divergence) takes the form of a Boltzmann distribution. Henceforth, minimizing the associated robust risk is equivalent with the optimization of the following objective:

$$\tilde{R}_n^{\text{KL}}(\theta) = \frac{\sum_{i=1}^n \ell(\xi_i;\theta)\exp(\ell(\xi_i;\theta)/\gamma^\star)}{\sum_{j=1}^n \exp(\ell(\xi_j;\theta)/\gamma^\star)}. \tag{20}$$

From an optimization standpoint this amounts to replacing the empirical distribution of the logged data with a Boltzmann adversary which re-weights samples in order to put more mass on hard examples (examples with high cost).

In what follows, we call KL-CRM the algorithm minimizing the objective (20) while treating $\gamma^\star$ as a hyper-parameter that controls the hardness of the re-weighting. A small value for $\gamma^\star$ will lead to a conservative behavior that will put more weight on actions with high propensity cost. In the limit when $\gamma^\star \to 0$, the robust risk only penalizes the action with highest propensity re-weighted cost. On the other end, a very large $\gamma^\star$ brings us back in the limit to the original CIPS estimator where the samples have equal weights. In a naive approach, this parameter can be determined through *cross-validation* and kept constant during the whole optimization procedure.

Lemma 5 goes further into the treatment of the optimal temperature parameter γ^* and provides an *adaptive* rule for updating it during the robust risk minimization procedure.

Lemma 5. *The value of the optimal temperature parameter γ^* can be approximated as follows:*

$$\gamma^* \approx \sqrt{\frac{V_n(\theta)}{2\epsilon}}. \tag{21}$$

Proof. To ease notations, we denote $Z \triangleq \ell(\xi, \theta)$. The log moment generating function :

$$\Phi : \alpha \to \log\mathbb{E}_{\hat{P}_n}\left[e^{Z\alpha}\right] \tag{22}$$

is well defined as $\hat{P}_n$ has finite support and Z is bounded a.s. It checks the following equalities:

$$\Phi(0) = 0$$
$$\Phi'(0) = \mathbb{E}_{\hat{P}_n}[Z]$$
$$\Phi''(0) = V_n(Z)$$

and a second-order Taylor expansion around 0 yields:

$$\Phi(\alpha) = \alpha\mathbb{E}_{\hat{P}_n}[Z] + \frac{\alpha^2}{2}V_n(Z) + o_0(\alpha^2)$$

With $\alpha = 1/\gamma$ and injecting this result in the r.h.s of Equation (18) yields:

$$\max_{KL(Q||\hat{P}_n)\leq\epsilon} \mathbb{E}_Q[\ell^M(\xi;\theta)] = \inf_{\gamma\geq 0}\Big\{\gamma\epsilon + \mathbb{E}_{\hat{P}_n}[Z] + \frac{V_n(Z)}{2\gamma} + o_\infty(1/\gamma)\Big\}$$

Solving (approximately) the r.h.s of the above equation yields as announced:

$$\gamma \simeq \sqrt{\frac{V_n(Z)}{2\epsilon}}$$

□

Algorithm 1: aKL-CRM

```
inputs : $\mathcal{H} = \{(x_1, y_1, p_1, c_1), \dots, (x_n, y_n, p_n, c_n)\}$,
         parametrized family of policies $\pi_\theta$
hyper-parameters: clipping constant $M$, uncertainty
         set size $\epsilon$
repeat
    compute the counterfactual costs
        $z_i \leftarrow c_i \min(M, \pi_\theta(y_i|x_i)/p_i)$ for $i = 1, \dots, n$
    compute the optimal temperature
        $\gamma^\star \leftarrow \sqrt{\sum_{j=1}^{n} (z_i - \bar{z})^2 / (2\epsilon)}$, where
        $\bar{z} = \sum_{j=1}^{n} z_i / n$
    compute the normalized costs $s_i \leftarrow e_i / \sum_{j=1}^{n} e_j$
        for $i = 1, \dots, n$, where $e_i = e^{z_i/\gamma^\star}$
    compute the re-weighted loss $L \leftarrow \sum_{i=1}^{n} z_i s_i$
    update $\theta$ by applying an L-BFGS step to the loss
        $L$
until convergence;
```

This results implies that γ^* should be updated concurrently to the parameter θ during the minimization of the robustified risk (20). This leads to an algorithm we call adaptive KL-CRM, or aKL-CRM. Pseudo-code for this algorithm is provided in Algorithm 1. As for POEM and KL-CRM, its hyper-parameter ϵ can be determined through cross-validation.

4 Experimental results

It is well known that experiments in the field of counterfactual reasoning are highly sensitive to differences in datasets and implementations. Consequently, to evaluate and compare the two algorithms we previously introduced to existing solutions, we rigorously follow the experimental procedure introduced in (Swaminathan and Joachims 2015a) and used in several other works - such as (Swaminathan and Joachims 2015b) since then. It relies on a supervised to unsupervised dataset conversion (Agarwal et al. 2014) to build bandit feedback from multi-label classification datasets. As in (Swaminathan and Joachims 2015a), we train exponential models

$$\pi_\theta(y|x) \propto \exp\left(\theta^T \phi(x, y)\right)$$

for the CRM problem and use the same datasets taken from the LibSVM repository. For reproducibility purposes, we used the code provided by its authors [1] for all our experiments.

4.1 Methodology

For any multi-label classification tasks, let us note x the input features and $y^\star \in \{0,1\}^q$ the labels. The full supervised dataset is denoted $\mathcal{D}^\star \triangleq \{(x_1, y_1^\star), \dots, (x_N, y_N^\star)\}$, and is split into three parts: $\mathcal{D}^\star_{\text{train}}$, $\mathcal{D}^\star_{\text{valid}}$, $\mathcal{D}^\star_{\text{test}}$. For every of the four dataset we consider (Scene, Yeast, RCV1-Topics and TMC2009), the split of the training dataset is done as follows: 75% goes to $\mathcal{D}^\star_{\text{train}}$ and 25% to $\mathcal{D}^\star_{\text{valid}}$. The test dataset $\mathcal{D}_{\text{test}}$ is provided by the original dataset. As in (Swaminathan and Joachims 2015a), we use joint features maps $\phi(x, y) = x \otimes y$ and train a Conditional Random Field (Lafferty, McCallum, and Pereira 2001) (CRF) on a fraction (5%, randomly constituted) of $\mathcal{D}^\star_{\text{train}}$. This CRF has access to the full supervised feedback and plays the role of the logging policy π_0. That is, for every $x_i \in \mathcal{D}^\star$, a label prediction y_i is sampled from the CRF with probability p_i. The quality of this prediction is measured through the Hamming loss: $c_i = \sum_{l=1}^{q} |y_l - y_l^\star|$. The logged bandit dataset is consequently generated by running this policy through $\mathcal{D}^\star_{\text{train}}$ for $\Delta = 4$ times (Δ is the *replay count*). After training, the performances of the different policies π are reported as their expected Hamming loss on the held-out set $\mathcal{D}^\star_{\text{test}}$. Every experiment is run 20 times with a different random seed (which controls the random training fraction for the logging policy and the creation of the bandit dataset).

	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 1: 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.

For each dataset we compare our algorithm with the naive CIPS estimator and the POEM. For all four algorithms (CIPS, POEM, KL-CRM , aKL-CRM), the numerical optimization routine is deferred to the L-BFGS algorithm. As in (Swaminathan and Joachims 2015a), the clipping constant M is always set to the ratio of the 90%ile to the 10%ile of the propensity scores observed in logs $\mathcal{H}$. Other hyper-parameters are selected by cross-validation on $\mathcal{D}^\star_{\text{valid}}$ with the unbiased counterfactual estimator (3). In the experimental results, we also report the performance of the logging policy π_0 on the test set as an indicative baseline measure, and the performance of a skyline CRF trained on the whole supervised dataset, despite of its unfair advantage.

4.2 Results

Table 1 reports the expected Hamming loss of the policies obtain with different algorithms on the Scene, Yeast, RCV1-Topics and TMC2007 dataset, averaged on 20 random seeds. The results reported for the baselines are coherent with (Swaminathan and Joachims 2015a). On each dataset, aKL-CRM comes out at one of the best algorithm (according to a one-tailed paired difference t-test at significance level 0.05) and outperforming the POEM baseline on three out of four datasets. The results for KL-CRM are more mitigated: it outperforms POEM on two datasets, but shows weaker performance on the two others clearly stressing the efficiency of an adaptive temperature parameter.

[1] http://www.cs.cornell.edu/~adith/POEM/index.html

	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 2: 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.

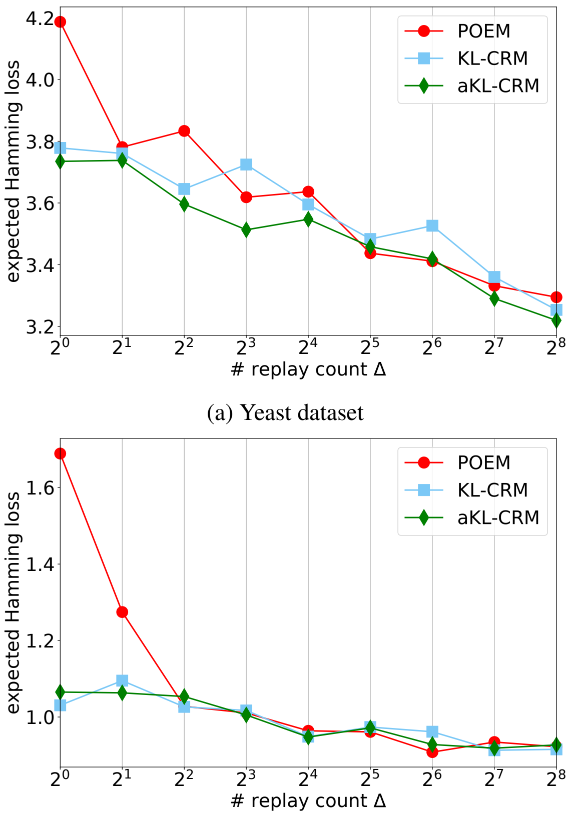
As in (Swaminathan and Joachims 2015a), we can further evaluate the quality of the learned policies by evaluating the Hamming loss of their *greedy* version (selecting only the arm that is attributed the most probability mass by the policy). This comes at less expense that sampling from the policy as this does not require to compute the normalizing constant in (6). These results are reported in Table 2, and are consistent with the conclusions of Table 1. One can note that the improvement brought by aKL-CRM over POEM is even sharper under this evaluation.

Another experiment carried in (Swaminathan and Joachims 2015a) focuses on the size of the bandit dataset $\mathcal{H}$. This quantity can be easily modulated by varying the *replay count* Δ - the number of times we cycle through $\mathcal{D}^{\star}_{\text{train}}$ to create the logged feedback $\mathcal{H}$. Figure 1 reports the expected Hamming loss of policies trained with the POEM, KL-CRM and aKL-CRM algorithms for different value of Δ, ranging from 1 to 256, and based on the Yeast and Scene datasets. Results are averaged over 10 independent runs. For large values of Δ (that is large bandit dataset) all algorithms seem to confound; this is to be expected as Lemma 2 states that for any coherent φ-divergences, all robust risks are asymptotically equivalent. It also stands out that for small replay counts (i.e the small data regime, a more realistic case), the KL-based algorithms outperform POEM.

5 Conclusion

We presented in this work a unified framework for counterfactual risk minimization based on the distributionally robust optimization of policies and motivated it by asymptotic guarantees available when the uncertainty measure is based on φ-divergences. We showed that this new framework generalizes existing solutions, like sample-variance penalized counterfactual risk minimization algorithms (Swaminathan and Joachims 2015a). Our work therefore opens a new avenue for reasoning about counterfactual optimization with logged bandit feedback as we showed that a KL-divergence based formulation of the counterfactual DRO problem can lead to tractable and efficient algorithms for the CRM problem, outperforming state-of-the-art results on a collection of datasets.

The authors of (Swaminathan and Joachims 2015a) also proposed a modification to the POEM algorithm that can be optimized at scale using stochastic gradient descent. Future work should therefore aim at developing stochastic optimization schemes for both KL-CRM and aKL-CRM so that they could handle large datasets. From the perspective of experimental evaluation, measuring the impact of the DRO formulation on the doubly robust (Dudík, Langford, and Li 2011) and the self-normalized (Swaminathan and Joachims 2015b) estimators would further validate its relevance for real world problems. Finally, a more theoretical line of work could focus on proving finite-samples performance certificate guarantees for distributionally robust estimators based on coherent φ-divergences, further motivating their use for counterfactual risk minimization.



(a) Yeast dataset

(b) Scene dataset

Figure 1: 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.

References

Agarwal, A.; Hsu, D.; Kale, S.; Langford, J.; Li, L.; and Schapire, R. 2014. Taming the Monster: A Fast and Simple Algorithm for Contextual Bandits. In *International Conference on Machine Learning*, 1638–1646.

Blanchet, J., and Murthy, K. 2019. Quantifying Distributional Model risk Via Optimal Transport. *Mathematics of Operations Research*.

Bottou, L.; Peters, J.; Candela, J. Q.; Charles, D. X.; Chickering, M.; Portugaly, E.; Ray, D.; Simard, P. Y.; and Snelson, E.

2013. Counterfactual Reasoning and Learning Systems: the example of Computational Advertising. *Journal of Machine Learning Research* 14(1):3207–3260.

Capen, E. C.; Clapp, R. V.; Campbell, W. M.; et al. 1971. Competitive Bidding in High-Risk Situations. *Journal of Petroleum Technology* 23(06):641–653.

Cortes, C.; Mansour, Y.; and Mohri, M. 2010. Learning Bounds for Importance Weighting. In *Advances in Neural Information Processing Systems*, 442–450.

Csiszár, I. 1967. Information-type measures of difference of probability distributions and indirect observation. *Studia Scientiarum Mathematicarum Hungarica* 2:229–318.

Duchi, J.; Glynn, P.; and Namkoong, H. 2016. Statistics of Robust Optimization: A Generalized Empirical Likelihood Approach.

Dudík, M.; Langford, J.; and Li, L. 2011. Doubly Robust Policy Evaluation and Learning. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, 1097–1104.

Esfahani, P. M., and Kuhn, D. 2018. Data-driven Distributionally Robust Optimization using the Wasserstein metric: Performance Guarantees and Tractable Reformulations. *Mathematical Programming* 171(1-2):115–166.

Esfahani et al., M. 2017. Data-driven distributionally robust optimization using the wasserstein metric: performance guarantees and tractable reformulations. *Mathematical Programming*.

Gotoh, J.; Kim, M. J.; and Lim, A. E. B. 2017. Calibration of distributionally robust empirical optimization models. *CoRR* abs/1711.06565.

Gotoh, J.; Kim, M. J.; and Lim, A. E. B. 2018. Robust empirical optimization is almost the same as mean-variance optimization. *Oper. Res. Lett.* 46(4):448–452.

Hu, Z., and Hong, L. J. 2013. Kullback-leibler divergence constrained distributionally robust optimization. *Available at Optimization Online*.

Hu, W.; Niu, G.; Sato, I.; and Sugiyama, M. 2018. Does Distributionally Robust Supervised Learning Give Robust Classifiers? In *International Conference on Machine Learning*, 2034–2042.

Ionides, E. L. 2008. Truncated Importance Sampling. *Journal of Computational and Graphical Statistics* 17(2):295–311.

Lafferty, J. D.; McCallum, A.; and Pereira, F. C. N. 2001. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, 282–289. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Li, L.; Chu, W.; Langford, J.; and Wang, X. 2011. Unbiased offline evaluation of contextual-bandit-based news article recommendation algorithms. In *Proceedings of the fourth ACM international conference on Web search and data mining*, 297–306. ACM.

Li, L.; Chen, S.; Kleban, J.; and Gupta, A. 2015. Counterfactual estimation and optimization of click metrics in search engines: A case study. In *Proceedings of the 24th International Conference on World Wide Web*, 929–934. ACM.

Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; and Vladu, A. 2017. Towards Deep Learning Models Resistant to Adversarial Attacks. *arXiv preprint arXiv:1706.06083*.

Maurer, A., and Pontil, M. 2009. Empirical Bernstein Bounds and Sample-Variance Penalization. In *COLT 2009 - The 22nd Conference on Learning Theory, Montreal, Quebec, Canada, June 18-21, 2009*.

Namkoong, H., and Duchi, J. C. 2017. Variance-based regularization with convex objectives. In *Advances in Neural Information Processing Systems*, 2971–2980.

Parys, B. P. G. V.; Esfahani, P. M.; and Kuhn, D. 2017. From Data to Decisions: Distributionally Robust Optimization is Optimal. *CoRR* abs/1704.04118.

Rockafellar, R. T. 2018. Risk and utility in the duality framework of convex analysis. *Springer Proceedings in Mathematics and Statistics*.

Rosenblum, P. R., and Rubin, D. B. 1983. The Central Role of the Propensity Score in Observational Studies for Causal Effect. *Biometrika* 70(1):41–55.

Sinha, A.; Namkoong, H.; and Duchi, J. C. 2017. Certifiable Distributional Robustness with Principled Adversarial Training. *CoRR* abs/1710.10571.

Smith, J. E., and Winkler, R. L. 2006. The Optimizer's Curse: Skepticism and Postdecision Surprise in Decision Analysis. *Management Science* 52(3):311–322.

Strehl, A.; Langford, J.; Li, L.; and Kakade, S. M. 2010. Learning from Logged Implicit Exploration Data. In *Advances in Neural Information Processing Systems*, 2217–2225.

Swaminathan, A., and Joachims, T. 2015a. Batch Learning from Logged Bandit Feedback through Counterfactual Risk Minimization. *Journal of Machine Learning Research* 16(1):1731–1755.

Swaminathan, A., and Joachims, T. 2015b. The Self-Normalized Estimator for Counterfactual Learning. In *Advances in Neural Information Processing Systems*, 3231–3239.

Thaler, R. 2012. *The winner's curse: Paradoxes and anomalies of economic life*. Simon and Schuster.

Van Parys, B. P.; Esfahani, P. M.; and Kuhn, D. 2017. From Data to Decisions: Distributionally Robust Optimization is Optimal. *arXiv preprint arXiv:1704.04118*.

Vapnik, V. 1992. Principles of Risk Minimization for Learning Theory. In *Advances in Neural Information Processing Systems*, 831–838.

Xiao, C.; Li, B.; Zhu, J.-Y.; He, W.; Liu, M.; and Song, D. X. 2018. Generating Adversarial Examples with Adversarial Networks. In *IJCAI*.