Improving Offline Contextual Bandits with Distributional Robustness

OTMANE SAKHI[*], LOUIS FAURY[*], and FLAVIAN VASILE, Criteo AI Lab, France.

This paper extends the Distributionally Robust Optimization (DRO) approach for offline contextual bandits laid out in [6]. Specifically, we leverage this framework to introduce a *convex* reformulation of the Counterfactual Risk Minimization principle introduced in [13]. Besides relying on convex programs, our approach is compatible with stochastic optimization, and can therefore be readily adapted to the large data regime. Our approach relies on the construction of asymptotic confidence intervals for offline contextual bandits through the DRO framework. By leveraging known asymptotic results of robust estimators, we also show how to automatically calibrate such confidence intervals, which in turn removes the burden of hyper-parameter selection for policy optimization. We present preliminary empirical results supporting the effectiveness of our approach.

ACM Reference Format:

Otmane Sakhi, Louis Faury, and Flavian Vasile. 2020. Improving Offline Contextual Bandits with Distributional Robustness. In *Proceedings of the ACM RecSys Workshop on Reinforcement Learning and Robust Estimators for Recommendation Systems (REVEAL '20).* ACM, New York, NY, USA, 10 pages.

1 INTRODUCTION

Contextual Bandits. The Contextual Bandit (CB) framework is a formalization of an important sequential decision making problem, with notorious applications in recommender systems [8, 17], mobile health [15] and clinical trials [18]. It describes a repeated game between a decision-maker and an environment. The latter sequentially reveal sets of available actions to the former, along with some additional side information (or *context*). Such additional information is assumed to carry informative signal about the intrinsic values (or *rewards*) of the actions. Informally, the goal of the decision-maker is to discover an efficient strategy (or *policy*) to select, given a context, a nearly optimal action.

Batch Learning from Bandit Feedback. The CB optimization literature can be divided into two streams. The first studies its *online* formulation, where the focus lies on the exploration/exploitation trade-off for *regret* minimization. This paper is concerned with the second, known as Batch Learning from Bandit Feedback [13] (BLBF) and arguably better suited for applications in real-life situations. The optimization is performed *offline* and based on historical data, typically obtained by logging the interactions between an older version of the current policy and the environment. The learning problem consists in leveraging this data (necessarily biased towards actions favored by the logged policy) to discover new strategies of greater performance.

Prior work and limitations. The first step in addressing the BLBF learning problem is to remove the intrinsic bias introduced by the logging policy [3]. This however can come at the price of building high-variance estimates [13] for the performance of the current policy, which in turns can lead to high *post-decision* regret. To address such challenges, the authors of [13] introduced the Counterfactual Risk Minimization (CRM) principle. It combines debiasing

[*] Equal contribution.

2020. Manuscript submitted to ACM

through importance re-weighting [13] with a modified policy-selection process that penalizes policies with high-variance estimates. Recently, [6] proposed a generalization of the CRM principle through the Distributionally Robust Optimization (DRO) framework. This led them to the development of a new BLBF algorithm, obtained through a specialization of this general framework . However, while the methods introduced in both [13] and [6] offer some desirable theoretical guarantees for off-line policy optimization, their respective implementation suffer from important caveats. Namely, they rely on optimizing *non-convex* objectives which is notably hard from a theoretical perspective. Further, these objective are not well-suited for stochastic optimization (useful when the logged data is large and cannot fit in memory), as obtaining unbiased stochastic gradients for these objective is not straight-forward. Finally, they rely on the selection of rather sensitive hyper-parameters, of which the (approximate) automatic calibration (through asymptotic arguments, for instance) is unknown.

Contributions. In this paper, we further investigate the DRO framework for offline CB introduced in [6]. This leads to a reformulation of the CRM principle that boils down to solving a convex problem. Further, we exploit some recent results from the stochastic optimization literature [10] to show how to efficiently solve this objective for large logged datasets. Our approach relies on the construction of asymptotic confidence intervals for offline CB through the DRO framework. Leveraging known asymptotic results for DRO we show how to automatically calibrate such confidence intervals, which in turns remove the need for hyper-parameter optimization for policy optimization. We validate our approach through extensive simulations on standard datasets for this task.

2 PRELIMINARIES

Notations. In the following φ is a real-valued convex function. The notation d_φ refers to the f-divergence associated to φ, and we denote $\varphi^\star(s) = \sup_{x\in\mathbb{R}}(xs - \varphi(x))$ the Fenchel conjugate of φ. We will write 1_n to be the n-dimensional vector which entries are all equal to $1/n$. For any positive integer m, Δ_m denotes the m-dimensional simplex.

Setting. In the following, we will use $x \in \mathcal{X}$ to denote a context and $a \in [K]$ an action, where K denotes the number of available actions. Given a context x, each action is associated with a cost[1] $c(x, a) \in [-1, 0]$, with the convention that better actions have smaller cost. The cost function c is unknown. A decision maker is represented by its policy π which maps each context $x \in \mathcal{X}$ to Δ_K. Assuming that the contexts are stochastic and follow a unknown distribution ν, we define the *risk* of the policy π as the expected cost one suffers when playing actions according to π:

$$\text{Risk}(\pi) = \mathbb{E}_{x\sim\nu, a\sim\pi(x)}\left[c(x,a)\right].$$

The learning problem is to find a policy π with smallest risk. In most real world problems, it is not reasonable to expect having the luxury of testing out several policies to compare their empirical risk and retain whichever policy has the smallest. This issue is usually circumvented by forecasting the risk of a given policy thanks to some existing interaction data. This is formalized through a *logging policy* π_0 which has already been deployed in the environment (e.g a previous version of a recommender system that the practitioner is trying to improved) for which we assume we have the history $\mathcal{H}_n = \{x_i, a_i \sim \pi_0(x_i), \pi_0(x_i, a_i), c(a_i, x_i)\}_{i\in[n]}$. Based on this data, one can build an unbiased (under mild assumptions) estimator of the risk of any policy π through the use of importance weights [11]:

$$\text{Risk}_n(\pi) := \frac{1}{n}\sum_{i=1}^{n} \omega_\pi(x_i, a_i) c(x_i, a_i) \quad \text{where} \quad w_\pi(x,a) := \pi(x,a)/\pi_0(x,a),$$

[1] We make this assumption for ease of exposition. It can be explicitly enforced by re-scaling the cost function.

commonly referred to as the IPS (Inverse Propensity Scoring) risk.

Counterfactual Risk Minimization. Unfortunately, the IPS estimator has potentially high variance, depending on the disparity between π and π_0 [13, Section 4]. Hence, directly sorting candidate policies thanks to their IPS risk is hazardous (as it boils down to comparing estimators with potentially high and different variances) and is known to be sub-optimal. To avoid this caveat, [13] proposed to add an *empirical variance* term to the IPS risk in order to penalize policies with high-variance estimates. Coined Counterfactual Risk Minimization (CRM), this principle suggest optimizing for the policy which minimizes:

$$\mathrm{Risk}_n^{\lambda}(\pi) = \mathrm{Risk}_n(\pi) + \lambda\sqrt{\frac{\widehat{\mathrm{Var}}_n(\pi)}{n}}\,, \tag{1}$$

where λ is a tunable hyper-parameter and $\widehat{\mathrm{Var}}_n(\pi) = \frac{1}{n-1}\sum_{i=1}^{n}\left(\omega_\pi(x_i,a_i)c(x_i,a_i) - \mathrm{Risk}_n(\pi)\right)^2$ is the empirical variance of $\mathrm{Risk}_n(\pi)$. This policy selection process is based on *variance-sensitive* confidence intervals for the true risk obtained via empirical Bernstein bounds [9].

Generalization through DRO. Based on a similar intuition, [6] recently introduced the idea of using Distributionally Robust Optimization (DRO) tools for this policy optimization problem. Formally, they showed that for a particular class of φ-divergence, the robust risk:

$$\mathrm{RobustRisk}_n^{\varphi}(\pi,\epsilon) := \sup_{q\in\Delta_n}\left\{\sum_{i=1}^{n} q_i\omega_\pi(x_i,a_i)c(x_i,a_i) \quad \text{s.t} \quad d_\varphi(q,1_n) \leq \epsilon\right\} \tag{2}$$

is a variance-sensitive (asymptotic) upper-bound for the true risk. It is therefore well-suited for the policy optimization task, and actually generalizes (in some sense) the CRM approach [6, Lemma 3]. For the KL-divergence, the robust risk has a closed-formed which can be directly minimized [6, Lemma 4]:

$$\mathrm{RobustRisk}_n^{\mathrm{KL}}(\pi,\epsilon) = \sum_{i=1}^{n}\frac{\exp(\omega_\pi(x_i,a_i)c(x_i,a_i)/\gamma)}{\sum_j \exp(\omega_\pi(x_j,a_j)c(x_j,a_j)/\gamma)}\omega_\pi(x_i,a_i)c(x_i,a_i)$$

with γ being a tunable hyper-parameter.

Limitations and contributions. The variance penalization of the IPS objective and its DRO generalization benefit from solid theoretical justifications, and result in better policy optimization algorithms. These algorithms, that were proven to outperform the simple IPS objective [6, 13] can still be improved as they suffer from important limitations: **(1)** Contrary to the IPS objective, which is linear (and therefore convex) in π, both the initial CRM objective (adding a square root variance penalization) and the DRO objective (as it was introduced in [6]) break this convexity. This results in ill-posed optimization programs, which potentially hinders the statistical benefits brought by such methods. **(2)** Another important limitation of such objectives is their scalability to large datasets. Both formulations are not well-adapted to stochastic gradient descent algorithms, as obtaining unbiased stochastic gradients of their related objectives is not-straightforward. For instance, the algorithm introduced in [6] works only in the batch setting as the adversary distribution needs all the data to be normalized. [13] suggested a relaxation of the CRM objective amenable to stochastic gradients, however only applicable in the case of exponential policies. Their approach consists in a majorization-minimization strategy, and still requires to go through the whole logged dataset once in a while. **(3)** These algorithms also come with hyper-parameters that need careful tuning as their choice drastically impact the performance of the obtained policy, making the whole optimization procedure even harder. [13] treats λ, the weight

Divergence	$\boldsymbol{\varphi(t)}$	$\boldsymbol{D_\varphi(q\|p)}$	$\boldsymbol{\varphi^*(s)}$
Chi-Square	$(t-1)^2$	$\sum_{i=1}^n \frac{(q_i-p_i)^2}{p_i}$	$\begin{cases} s+s^2/4 & s\geq -2 \\ -1 & s \leq -2 \end{cases}$
Kullback-Leibler	$t\log t - t + 1$	$\sum_{i=1}^n q_i \log(q_i/p_i)$	$e^s - 1$
Burg entropy	$-\log t + t - 1$	$\sum_{i=1}^n p_i \log(p_i/q_i)$	$-\log(1-s), s<1$
Hellinger distance	$(\sqrt{t}-1)^2$	$\sum_{i=1}^n \left(\sqrt{p_i}-\sqrt{q_i}\right)^2$	$\frac{s}{1-s}, s \leq 1$

Table 1. Some coherent φ-divergences and their characterizations.

of the variance penalty, as a hyper-parameter, while [6] treat ϵ, the maximum distance between the adversarial and the nominal distribution as a hyper-parameter as well. To sum-up, both algorithms are deemed rather impractical in real life, as they have a non convex loss surface to optimize, are not applicable to huge datasets and need consequent hyper-parameter tuning. This work tries to circumvent these limitations through a more careful treatment of the DRO formulation, leading to general algorithms that treat all these caveats in a well defined and unified framework.

3 POLICY EVALUATION AND OPTIMIZATION

3.1 Policy Evaluation: Confidence Intervals

In this section, we briefly review and discuss how the robust risk can lead to the construction of confidence intervals for the true risk. This is a crucial step for policy optimization, as the latter will consist in minimizing the high-probability upper-bound on the true risk provided by the policy evaluation procedure. Designing tight confidence interval for the risk can also be a goal in itself, in order to fully evaluate the potential benefits/risks of deploying a given policy (e.g provide an offline metric for A/B testing). In the following, we follow [6] and consider *coherent* φ-divergence - i.e we will assume that φ satisfies the conditions of Assumption 1 in [6]. We provide some examples of such functions (and their associated divergence measure) in Table 1.

Under such conditions, one can show that the robust risk accounts for the variance of the IPS risk estimator [6, Lemma 2]. Further, DRO can be further leveraged to build asymptotic confidence intervals for the robust-risk. Indeed, let us introduce the following problem, converse to (2):

$$\text{OptimisticRisk}_n^\varphi(\pi,\epsilon) := \inf_{q\in\Delta_n}\left\{\sum_{i=1}^n q_i\omega_\pi(x_i,a_i)c(x_i,a_i) \quad \text{s.t} \quad d_\varphi(q,1_n)\leq\epsilon\right\}. \tag{3}$$

We have the following result, extending Lemma 1 of [6] and easily extracted from [5].

Proposition 1. *[Asymptotic Confidence Interval] Let $\delta\in[0,1)$. For $\alpha\in(0,1)$ denote ρ_α the $(1-\alpha)$-quantile of the one-dimensional χ^2 distribution. Then:*

$$\lim_{n\to\infty}\mathbb{P}\left(\text{OptimisticRisk}_n^\varphi(\pi,\frac{\rho_\delta}{n})\leq \text{Risk}(\pi)\leq \text{RobustRisk}_n^\varphi(\pi,\frac{\rho_\delta}{n})\right)\geq 1-\delta\,. \tag{4}$$

This result states that the interval $[\text{OptimisticRisk}_n^\varphi(\pi,\epsilon),\text{RobustRisk}_n^\varphi(\pi,\epsilon)]$ is a asymptotic $(1-\delta)$ confidence interval for the true risk, when the size of the ambiguity-set ϵ is set to ρ_δ/n. We will show in Section 4.1 that despite being asymptotic, this interval is empirically *tight* and displays satisfying coverage, motivating its use in real-life applications. It turns out that the programs for computing the robust and optimistic risk (Equation (2) and (3), respectively) can be efficiently solved. We will only review here the computation of the robust risk, however a similar reasoning holds for the optimistic risk. Notice that the objective in Equation (2) is linear in the variable q, which acts as a re-weighting for the counterfactual costs. Further, the constraint set $\{q \in \Delta_n \mid d_\varphi(q, 1_n) \leq \epsilon\}$ is convex. The program is therefore *convex* and can henceforth be solved efficiently. In this paper, we will rely on its dual formulation, which is easily solvable and well-adapted to the stochastic setting. Formally, we rely on the following result to characterize the robust risk.

Lemma 1. *[Dual program for the robust risk] Let:*

$$g_\pi(\beta, \gamma) = \beta + \gamma\epsilon + \frac{1}{n}\sum_{s=1}^{n}(\gamma\varphi)^\star\left(\omega_\pi(x_i, a_i)c(a_i, x_i) - \beta\right) , \tag{5}$$

where $(\gamma\varphi)^\star(s) = \gamma\varphi^\star(s/\gamma)$*, with the convention that* $(0\varphi)^\star(s) = +\infty$ *is* $s > 0$ *and* 0 *otherwise. The function* $(\pi, \beta, \gamma) \to g_\pi(\beta, \gamma)$ *is convex and:*

$$\text{RobustRisk}_n^\varphi(\pi, \epsilon) = \inf_{\beta, \gamma \geq 0} g_\pi(\beta, \gamma) \,. \tag{\textbf{DRO-PE}}$$

This robust program characterization can be extracted from more general results - see for instance [2, Section 4]. We provide a detailed proof in Appendix A for the sake of completeness. In a few words, Lemma 1 states the the robust risk can be efficiently computed by solving a two-dimensional convex program. When n is reasonably small (in other words, when the dataset $\mathcal{H}_n$ fits in memory), coordinate descent (with exact line search) or two-dimensional bisection provide efficient, principled tools for computing the robust risk. The program (**DRO-PE**) is also well-suited for the large-data regime (*e.g* large n) as it naturally adapts to stochastic optimization. Indeed, the function g (Equation (5)) is composite and unbiased gradients of this objective are easily obtainable. Stochastic gradient descent methods (see [12] for a modern overview) therefore provide efficient and flexible solutions for this problem (up to some mild modifications to account for the fact that the g is not smooth for γ in a neighborhood of 0).

To sum-up, we showed here how the DRO method could be used to build confidence intervals for the true risk, by simply relying on solving convex programs. This confidence intervals are however asymptotic; we will show in Section 4 that, still, they provide sufficient coverage, while being much tighter than their finite-time counterparts.

3.2 Policy Optimization: Towards a Convex Objective

3.2.1 General principle. The CRM principle casts policy optimization in a theoretically sound framework, and interestingly enough is closely related to the robust risk defined throughout the paper. Relying on [5], [6] namely showed that the robust risk provides an asymptotic approximation to the variance regularized empirical risk. The original CRM principle for policy optimization (Equation (1)) can therefore be rethought as the minimization of a φ-robust risk. Using the dual formulation of the robust risk given in Equation (**DRO-PE**), the policy optimization objective becomes:

$$\inf_\pi \text{RobustRisk}_n^\varphi(\pi, \epsilon) = \inf_{\pi, \beta, \gamma \geq 0} g_\pi(\beta, \gamma) \,. \tag{\textbf{DRO-PO}}$$

Note that as a consequence of Lemma 1, this objective is *convex*. It can therefore be minimized in principled ways, while enjoying similar guarantees as the original CRM objective. Intuitively, one can expect such an important transformation of the optimization properties of the policy improvement objective to lead to greater practical performances. The most natural way to solve the policy improvement objective (**DRO-PO**) is through plain gradient descent. Indeed, a valid strategy consists in feeding gradients of the function $(\beta, \gamma, \pi) \to g_\pi(\beta, \gamma)$ to a gradient optimizer. In our experiments, we found that applying the L-BFGS solver to this dual program to work best. As for the policy evaluation, the policy optimization objective (**DRO-PO**) is particularly adapted when the historic data is large (i.e $n \gg 1$) and only stochastic gradients can be obtained. For stochastic optimization, stochastic gradients methods can encounter some issues (linked to the possible unboundedness of the gradients) which can be alleviate thanks to specialized methods [10].

3.2.2 Extensions. In the following, we discuss how different estimators can be used and robustified in the same way as the IPS, for improved performances and without sacrificing convexity.

Variance Reduction. The methods presented so far rely on vanilla IPS. It is well known that this estimator suffers from large variance which can lead to poor performances - whatever the policy optimization algorithm used. Fortunately, our method easily extend to other estimators, so long that they remain convex in π. Still, it would be useful to extend our method to estimators that actively reduce variance. A candidate for this task is the self-normalized importance sampling estimator of [14]. This estimator is unfortunately not convex in π, which goes against the efforts undertaken in this paper to maintain well-behaved optimization tasks. We provide here an alternative which uses a simple additive control variate (instead of a multiplicative one). Formally, we rely on the following estimator:

$$\mathrm{Risk}_{n,\rho}(\pi) = \frac{1}{n}\sum_{i=1}^{n} \left(c(x_i, a_i) - \rho\right) \omega_\pi(x_i, a_i) + \rho\,.$$

A robust version of this estimator easily follows, and enjoys the same convex properties of the IPS robust risk. The variance-reduction property of the additive control variate is presented in the following Lemma.

Lemma 2. *[Propensity weights as an additive control variate] For all ρ, $\mathrm{Risk}_{n,\rho}(\pi)$ is an unbiased estimator of*$\mathrm{Risk}(\pi)$*, achieving a better variance than naive IPS whenever $0 \leq \rho \leq 2\frac{Cov(\ell_\pi,\omega_\pi)}{\mathbf{V}(\omega_\pi)}$. In addition, if the cost is independent of the propensity weights, we obtain $\rho^* = \mathbb{E}[c]$.*

In practice, we do not know how to derive ρ^* analytically, however one can directly use the cost's empirical mean under π_0.

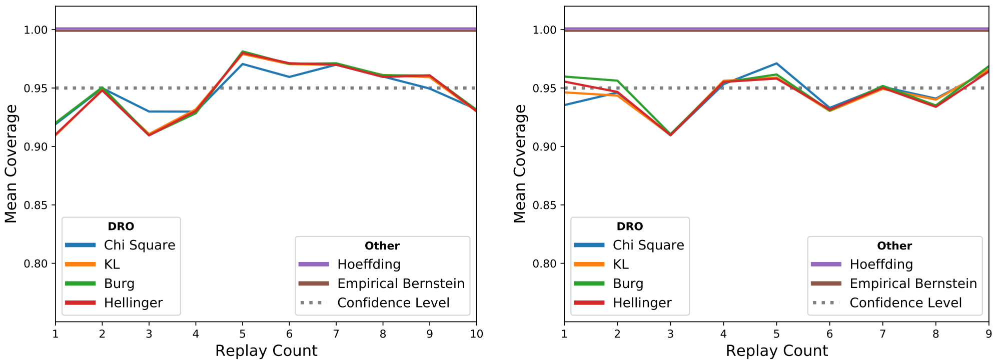
Parametric Policies. In practice, the actions/contexts space is extremely large and directly optimizing the objective with respect to the policy (as a $\mathbb{R}^{|\mathcal{X}|\times K}$ matrix) is unreasonable. In such cases, policies are parametrized to drastically reduce the complexity of the problem. This usually breaks convexity, even in the simplest case of log-linear policies - that is, policies of the form $\pi_\theta(a|x) \propto \exp(\theta^T f(x, a))$ for $f(x, a)$ a given joint feature map. In this case, the objective becomes a negative sum of log-concave functions resulting in a non-convex optimization surface. Following [7], one can bypass this non-convexity by constructing a *tight* convex upper bound of the original objective.

Lemma 3. *[Convex upper-bound for log-concave policies] Let π_θ be a log-concave (w.r.t θ) policy. For a given θ_0, let:*

$$\mathrm{Risk}_n^{up}(\pi_\theta) = \frac{1}{n}\sum_{i=1}^{n} \frac{\pi_{\theta_0}(a_i, x_i)}{\pi_0(a_i, x_i)}\left(1 + \log\left[\frac{\pi_\theta(a_i, x_i)}{\pi_{\theta_0}(a_i, x_i)}\right]\right)c(a_i, x_i)\,.$$

$\mathrm{Risk}_n^{up}(\pi_\theta)$ *is a* convex *upper bound of the IPS risk. The closer θ_0 to θ, the tighter the upper bound, with equality at $\theta_0 = \theta$.*

We can use Lemma 3 to obtain a proxy of our initial objective, building on an iterative procedure that only uses convex losses throughout the whole optimization process. Once again, we can build a robust version of this estimator which can be efficiently optimized. Note that here, the robust estimator will be convex w.r.t the parametrization θ as soon as the policy in log-concave.



(a) Empirical coverage of the true risk for different confidence intervals on the Yeast dataset.

(b) Empirical coverage of the true risk for different confidence intervals on the Scene dataset.

Fig. 1. Finite-time evaluation (coverage) for asymptotic DRO-based confidence intervals.

4 PRELIMINARY RESULTS

We here describe some preliminary experimental results, backing up the idea that an improved optimization landscape for policy optimization naturally leads to improved practical performances. We work with the four φ-divergence presented in Table 1. We employ the classical supervised to bandit conversion [1]. Formally, denote $x \in \mathcal{X}$ a given input vector and $y \in \{0,1\}^L$ its label, and $\mathcal{D}^\star = \{(x_1, t_1), \dots, (x_m, t_m)\}$ a given multi-label dataset. We create the logging policy by training it on a fraction of $\mathcal{D}^\star$. We then create the historic data $\mathcal{H}_n$ by going repeating P times the following procedure: for every (x_i, t_i) in the supervised dataset, sample $a_i \sim \pi_0(x)$ and log the cost $c(a_i, x_i) = \|a_i - t_i\|_1$. Following [13], we call P the *replay count*.

4.1 DRO Confidence Intervals

We start this experimental section by performing a *sanity check* on the (asymptotic) confidence intervals that we based our policy optimization method on. Formally, we evaluate the finite-time validity of Equation (4). Being asymptotic, we can safely expect DRO-based confidence intervals to be smaller than their finite-time counterparts - i.e confidence intervals based on Hoeffding or empirical Bernstein tail inequalities (see [16] for their application to policy evaluation). We however wish to check that they provide reasonable coverage in non-asymptotic regimes. To do so, we train a policy π on a random subset of $\mathcal{D}^\star$ and evaluate the empirical mean coverage and width of DRO-based intervals. In Figure 1, we present such results on two datasets: Yeast and Scene, taken from the LibSVM repository and standard for the policy optimization task [13]. The empirical coverageis reported for increasing values of the replay count P, or equivalently for increasing values of the historic data size n. The failure level is set to $\delta = 0.95$ in all experiments. We observe that the DRO- ased confidence interval provide almost exact $(1 - \delta)$ coverage, and it is therefore safe to use them even in the small data regime. As a side comment, we observe that all four φ-divergence lead to very similar results. Finally, we check experimentally that as expected, the asymptotic DRO-based confidence intervals are by orders of magnitude smaller than finite-time ones.

Algorithm	Risk	Greedy-Risk
POEM-$\mathcal{B}$	0.93 (0.06)	0.91 (0.06)
DRO-$\mathcal{B}$-χ^2	**0.89** (0.06)	0.87 (0.06)
DRO-$\mathcal{B}$-KL	0.90 (0.06)	0.88 (0.06)
DRO-$\mathcal{B}$-Burg	1.06 (0.06)	**0.85** (0.05)
DRO-$\mathcal{B}$-Hellinger	1.06 (0.06)	0.85 (0.05)
POEM-$\mathcal{S}$	**1.0** (0.05)	**0.97** (0.05)
DRO-$\mathcal{S}$-χ^2	1.05 (0.06)	1.02 (0.06)
DRO-$\mathcal{S}$-KL	1.06 (0.07)	1.04 (0.07)
DRO-$\mathcal{S}$-Burg	1.12 (0.05)	1.08 (0.05)
DRO-$\mathcal{S}$-Hellinger	1.3 (0.08)	1.18 (0.07)

(a) Scene dataset. The quantity in parenthesis represent one standard deviation of the aggregated results.

Algorithm	Risk	Greedy-Risk
POEM-$\mathcal{B}$	5.15 (0.07)	4.34 (0.13)
DRO-$\mathcal{B}$-χ^2	5.21 (0.05)	4.38 (0.12)
DRO-$\mathcal{B}$-KL	5.32 (0.04)	5.29 (0.11)
DRO-$\mathcal{B}$-Burg	**4.77** (0.07)	**3.74** (0.13)
DRO-$\mathcal{B}$-Hellinger	**4.77** (0.07)	**3.74** (0.13)
POEM-$\mathcal{S}$	**5.16** (0.05)	**4.62** (0.1)
DRO-$\mathcal{S}$-χ^2	5.17 (0.06)	4.71 (0.11)
DRO-$\mathcal{S}$-KL	5.17 (0.06)	4.72 (0.12)
DRO-$\mathcal{S}$-Burg	5.17 (0.06)	4.72 (0.1)
DRO-$\mathcal{S}$-Hellinger	5.27 (0.06)	4.71 (0.1)

(b) Yeast dataset. The quantity in parenthesis represent one standard deviation of the aggregated results.

Fig. 2. Policy optimization results.

4.2 Policy Optimization

We report here some preliminary results for the policy optimization, for which we strictly follow the experimental procedure of [13]. The supervised dataset $\mathcal{D}^\star$ is split into three parts (train, validation and test). A logging policy is train on a random fraction (= 0.1) of $\mathcal{D}_*$ and used to collect the history $\mathcal{H}_n$ by running it through the training data $P = 4$ times. For DRO-based algorithms, the validation set is not used, since no hyper-parameter needs to be tuned (we use the value recommended by the asymptotic analysis for ε, with a fixed confidence level at $\delta = 0.05$). For POEM and its stochastic approximation, the parameter λ is selected by cross-validation on the validation data. We report in Figure 2 the risks of the policy (and their greedy versions) returned by the different algorithms. As in [13], all policies are parametrized linearly, with a softmax output activation layer. We present results for both batch algorithms ($\mathcal{B}$) and their stochastic versions ($\mathcal{S}$). Results are average over 20 random repetitions.

5 DISCUSSION

For batch algorithms, one can notice that DRO-based methods provide either similar or better empirical results than POEM on both considered datasets, while being hyper-parameter free (which is not the case of POEM). On the Yeast dataset, the improvement is quite significative for two of the four φ-divergence (Burg and Hellinger). On the negative side, it seems there is no consistency in the relative performance of the different divergences. This is quite troublesome in practice, as to the best of our knowledge there is no obvious nor preferable choice of divergences given a dataset. A solution to this problem is probably to cross-validate this choice, potentially over a continuous parametrization of the divergence considered here (such as the parameter of a Cressie-Read divergence). Finally, we note that POEM-$\mathcal{S}$ dominates among the stochastic algorithm considered. This is however to be nuanced, as this algorithm still needs to load in memory the entire dataset at every epoch (*e.g.* every time an upper-bound on the true objective is constructed). This is not the case for DRO-based algorithms. We also believe that the nonetheless good performances reported here for stochastic DRO algorithms will turn to be decisive when considering more complex policies (*e.g* parametrized by a neural network, where POEM-$\mathcal{S}$ have been reported to fail). We plan on investigating this in future work, along with the performance of other robustified estimators (cf. Lemma 2 and 3.)

REFERENCES

[1] Alekh Agarwal, Daniel Hsu, Satyen Kale, John Langford, Lihong Li, and Robert Schapire. 2014. Taming the Monster: A Fast and Simple Algorithm for Contextual Bandits. In *International Conference on Machine Learning*. 1638–1646.

[2] Aharon Ben-Tal, Dick Den Hertog, Anja De Waegenaere, Bertrand Melenberg, and Gijs Rennen. 2013. Robust solutions of optimization problems affected by uncertain probabilities. *Management Science* 59, 2 (2013), 341–357.

[3] Léon Bottou, Jonas Peters, Joaquin Quiñonero-Candela, Denis X. Charles, D. Max Chickering, Elon Portugaly, Dipankar Ray, Patrice Simard, and Ed Snelson. 2013. Counterfactual Reasoning and Learning Systems: The Example of Computational Advertising. *Journal of Machine Learning Research* 14, 65 (2013), 3207–3260. http://jmlr.org/papers/v14/bottou13a.html

[4] Patrick L Combettes. 2018. Perspective Functions: Properties, Constructions, and Examples. *Set-Valued and Variational Analysis* 26, 2 (2018), 247–264.

[5] John Duchi, Peter Glynn, and Hongseok Namkoong. 2016. Statistics of Robust Optimization: A Generalized Empirical Likelihood Approach. *arXiv preprint arXiv:1610.03425* (2016).

[6] Louis Faury, Ugo Tanielan, Flavian Vasile, Elena Smirnova, and Elvis Dohmatob. 2020. Distributionally Robust Counterfactual Risk Minimization. In *Thirty-Fourth AAAI Conference on Artificial Intelligence*.

[7] Nicolas Le Roux. 2016. Tighter bounds lead to improved classifiers. *CoRR* abs/1606.09202 (2016). arXiv:1606.09202 http://arxiv.org/abs/1606.09202

[8] Lihong Li, Wei Chu, John Langford, and Robert E Schapire. 2010. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World Wide Web*. 661–670.

[9] Andreas Maurer and Massimiliano Pontil. 2009. Empirical Bernstein Bounds and Sample Variance Penalization. In *Proceedings of the 22nd Conference on Learning Theory*.

[10] Hongseok Namkoong and John C Duchi. 2016. Stochastic Gradient Methods for Distributionally Robust Optimization with f-divergences. In *Advances in Neural Information Processing Systems*. 2208–2216.

[11] Paul R Rosenbaum and Donald B Rubin. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika* 70, 1 (1983), 41–55.

[12] Sebastian Ruder. 2016. An overview of gradient descent optimization algorithms. *arXiv preprint arXiv:1609.04747* (2016).

[13] Adith Swaminathan and Thorsten Joachims. 2015. Batch Learning from Logged Bandit Feedback through Counterfactual Risk Minimization. *Journal of Machine Learning Research* 16, 52 (2015), 1731–1755. http://jmlr.org/papers/v16/swaminathan15a.html

[14] Adith Swaminathan and Thorsten Joachims. 2015. The Self-Normalized Estimator for Counterfactual Learning. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2 (NIPS'15)*. MIT Press, Cambridge, MA, USA, 3231–3239.

[15] Ambuj Tewari and Susan A Murphy. 2017. From Ads to Interventions: Contextual Bandits in Mobile Health. In *Mobile Health*. Springer, 495–517.

[16] Philip S Thomas, Georgios Theocharous, and Mohammad Ghavamzadeh. 2015. High-Confidence Off-Policy Evaluation. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*.

[17] Michal Valko, Rémi Munos, Branislav Kveton, and Tomáš Kocák. 2014. Spectral Bandits for Smooth Graph Functions. In *International Conference on Machine Learning*. 46–54.

[18] Sofía S Villar, Jack Bowden, and James Wason. 2015. Multi-armed bandit models for the optimal design of clinical trials: benefits and challenges. *Statistical science: a review journal of the Institute of Mathematical Statistics* 30, 2 (2015), 199.

A PROOF OF LEMMA 1

Lemma 1. *[Dual program for the robust risk] Let:*

$$g_\pi(\beta,\gamma) = \beta + \gamma\epsilon + \frac{1}{n}\sum_{s=1}^{n}(\gamma\varphi)^\star\left(\omega_\pi(x_i,a_i)c(a_i,x_i) - \beta\right) , \tag{5}$$

where $(\gamma\varphi)^\star(s) = \gamma\varphi^\star(s/\gamma)$, with the convention that $(0\varphi)^\star(s) = +\infty$ is $s > 0$ and 0 otherwise. The function $(\pi,\beta,\gamma) \to g_\pi(\beta,\gamma)$ is convex and:

$$\text{RobustRisk}_n^\varphi(\pi,\epsilon) = \inf_{\beta,\gamma\geq 0} g_\pi(\beta,\gamma) \,. \tag{\textbf{DRO-PE}}$$

Proof. Recall the definition of the robust risk:

$$\text{RobustRisk}_n^\varphi(\pi,\epsilon) := \sup_{q\in\Delta_n}\left\{\sum_{i=1}^{n} q_i\omega_\pi(x_i,a_i)c(x_i,a_i) \quad \text{s.t} \quad d_\varphi(q,1_n) \leq \epsilon\right\} \tag{\textbf{P}}$$

where:

$$\begin{cases} \Delta_n = \left\{ p \in \mathbb{R}_n^+ \middle| \sum_{i=1}^n p_i = 1 \right\} \\ 1_n = \frac{1}{n}(1 \ldots 1)^\mathsf{T} \in \mathbb{R}^n \\ d_\varphi(q, p) = \sum_{i=1}^n p_i \varphi\left(\frac{q_i}{p_i}\right) \quad \forall q \ll p \in \Delta_n \end{cases}$$

Note that the program (**P**) optimizes a linear objective under convex constraints (since φ is convex). Further, when $\epsilon > 0$, the candidate $q = 1_n$ is *strictly* feasible. Therefore, Slater's condition holds and (**P**) enjoys strong duality. Writing down its Lagrangian, we obtain the following equivalence:

$$\begin{aligned} \text{RobustRisk}_n^\varphi(\pi, \epsilon) &= \sup_{q \geq 0} \inf_{\beta, \gamma \geq 0} \sum_{i=1}^n q_i \omega_\pi(x_i, a_i) c(x_i, a_i) + \beta\left(1 - \sum_{i=1}^n q_i\right) + \gamma\left(\epsilon - \frac{1}{n}\sum_{i=1}^n \varphi(nq_i)\right) \\ &= \inf_{\beta, \gamma \geq 0} \sup_{q \geq 0} \sum_{i=1}^n q_i \omega_\pi(x_i, a_i) c(x_i, a_i) + \beta\left(1 - \sum_{i=1}^n q_i\right) + \gamma\left(\epsilon - \frac{1}{n}\sum_{i=1}^n \varphi(nq_i)\right) \\ &= \inf_{\beta, \gamma \geq 0} \beta + \gamma\epsilon + \frac{1}{n}\sum_{i=1}^n \sup_{q_i \geq 0} \left\{(nq_i)\omega_\pi(x_i, a_i)c(x_i, a_i) - \gamma\varphi(nq_i)\right\} \end{aligned} \tag{6}$$

where the first equality is a consequence of strong duality, and the second is obtained through simple re-arranging. If $\gamma \neq 0$, easy computations lead to:

$$\begin{aligned} \text{RobustRisk}_n^\varphi(\pi, \epsilon) &= \inf_{\beta, \gamma \geq 0} \beta + \gamma\epsilon + \frac{\gamma}{n}\sum_{i=1}^n \sup_{q_i \geq 0} \left\{(nq_i)\frac{\omega_\pi(x_i, a_i)c(x_i, a_i)}{\gamma} - \varphi(nq_i)\right\} \\ &= \inf_{\beta, \gamma \geq 0} \beta + \gamma\epsilon + \frac{\gamma}{n}\sum_{i=1}^n \varphi^\star\left(\frac{\omega_\pi(x_i, a_i)c(x_i, a_i)}{\gamma}\right) \end{aligned}$$

by using the definition of $\varphi^\star$. The limit conditions announced in the Lemma are easily checked by computing the dual function when $\gamma = 0$. We therefore obtain the equality announced by using the definition of g_π:

$$\text{RobustRisk}_n^\varphi(\pi, \epsilon) = \inf_{\beta, \gamma \geq 0} g_\pi(\beta, \gamma)$$

The convexity of g_π can be obtained two ways; (1) by noticing that g_π is obtained through convexity-transforming transformations of a *perspective* function [4], or (2) by noticing thanks to Equation (6) that:

$$(\pi, \beta, \gamma) \to \sum_{i=1}^n \sup_{q_i \geq 0} \left\{(nq_i)\omega_\pi(x_i, a_i)c(x_i, a_i) - \gamma\varphi(nq_i)\right\}$$

is convex as a sum of supremum of linear (and hence convex) functions. □