Multi-objective Bandits: Optimizing the Generalized Gini Index

Róbert Busa-Fekete [1] **Balázs Szörényi** [2 3] **Paul Weng** [4 5] **Shie Mannor** [3]

Abstract

We study the multi-armed bandit (MAB) problem where the agent receives a vectorial feedback that encodes many possibly competing objectives to be optimized. The goal of the agent is to find a policy, which can optimize these objectives simultaneously in a fair way. This multi-objective online optimization problem is formalized by using the Generalized Gini Index (GGI) aggregation function. We propose an online gradient descent algorithm which exploits the convexity of the GGI aggregation function, and controls the exploration in a careful way achieving a distribution-free regret $\tilde{O}(T^{-1/2})$ with high probability. We test our algorithm on synthetic data as well as on an electric battery control problem where the goal is to trade off the use of the different cells of a battery in order to balance their respective degradation rates.

1. Introduction

The multi-armed bandit (MAB) problem (or bandit problem) refers to an iterative decision making problem in which an agent repeatedly chooses among K options, metaphorically corresponding to pulling one of K arms of a bandit machine. In each round, the agent receives a random payoff, which is a reward or a cost that depends on the arm being selected. The agent's goal is to optimize an evaluation metric, e.g., the *error rate* (expected percentage of times a suboptimal arm is played) or the *cumulative regret* (difference between the sum of payoffs obtained and the (expected) payoffs that could have been obtained by selecting the best arm in each round). In the *stochastic* multi-armed bandit setup, the payoffs are assumed to obey fixed distributions that can vary with the arms but do not change with time. To achieve the desired goal, the agent has to tackle the classical exploration/exploitation dilemma: It has to properly balance the pulling of arms that were found to yield low costs in earlier rounds and the selection of arms that have not yet been tested often enough (Auer et al., 2002a; Lai & Robbins, 1985).

The bandit setup has become the standard modeling framework for many practical applications, such as online advertisement (Slivkins, 2014) or medical treatment design (Press, 2009) to name a few. In these tasks, the feedback is formulated as a single real value. However many real-world online learning problems are rather multi-objective, i.e., the feedback consists of a vectorial payoffs. For example, in our motivating example, namely an electric battery control problem, the learner tries to discover a "best" battery controller, which balances the degradation rates of the battery cells (i.e., components of a battery), among a set of controllers while facing a stochastic power demand. Besides, there are several studies published recently that consider multi-objective sequential decision problem under uncertainty (Drugan & Nowé, 2013; Roijers et al., 2013; Mahdavi et al., 2013).

In this paper, we formalize the multi-objective multi-armed bandit setting where the feedback received by the agent is in the form of a D-dimensional cost vector. The goal here is to be both efficient, i.e., minimize the cumulative cost for each objective, and fair, i.e., balance the different objectives. One natural way to ensure this is to try to find a cost vector on the Pareto front that is closest to the origin or to some other ideal point. A generalization of this approach (when using the infinite norm) is the Generalized Gini Index (GGI), a well-known inequality measure in economics (Weymark, 1981).

GGI is convex, which suggests applying the Online Convex Optimization (OCO) techniques (Hazan, 2016; Shalev-Shwartz, 2012). However, a direct application of this technique may fail to optimize GGI under noise, because the objective can be only observed with a bias that is induced by the randomness of the cost vectors and by the fact that the performance is measured by the function value of the average cost instead of the average of the costs' function value. The solution we propose is an online learning algorithm which is based on Online Gradient Descent (OGD) (Zinkevich, 2003; Hazan, 2016) with additional exploration that enables us to control the bias of the objective function. We also show that its regret is almost optimal: up to a logarithmic factor,

[*]Equal contribution [1]Yahoo Research, New York, NY, USA [2]Research Group on AI, Hungarian Acad. Sci. and Univ. of Szeged, Szeged, Hungary [3]Technion Institute of Technology, Haifa, Israel [4]SYSU-CMU JIE, SEIT, SYSU, Guangzhou, P.R. China [5]SYSU-CMU JRI, Shunde, P.R. China. Correspondence to: Paul Weng <paul@weng.fr>.

Proceedings of the 34^{th} International Conference on Machine Learning, Sydney, Australia, PMLR 70, 2017.

it matches the distribution-free lower bound of the stochastic bandit problem (Auer et al., 2002b), which naturally applies to our setup when the feedback is one-dimensional.

The paper is organized as follows: after we introduce the formal learning setup, we briefly recall the necessary notions from multi-objective optimization in Section 3. Next, in Section 4, GGI is introduced and some of its properties are described. In Sections 5 and 6, we present how to compute the optimal policy for GGI, and define the regret notion. Section 7 contains our main results where we define our OGD-based algorithm and analyze its regret. In Section 8, we test our algorithm and demonstrate its versatility in synthetic and real-world battery-control experiments. In Section 9, we provide a survey of related work, and finally conclude the paper in Section 10.

2. Formal setup

The multi-armed or K-armed bandit problem is specified by real-valued random variables $X_1, \dots, X_K$ associated, respectively, with K arms (that we simply identify by the numbers $1, \dots, K$). In each time step t, the online learner selects one and obtains a random sample of the corresponding distributions. These samples, which are called costs, are assumed to be independent of all previous actions and costs.[1] The goal of the learner can be defined in different ways, such as minimizing the sum of costs over time (Lai & Robbins, 1985; Auer et al., 2002a).

In the *multi-objective* multi-armed bandit (MO-MAB) problem, costs are not scalar real values, but real vectors. More specifically, a D-objective K-armed bandit problem ($D \geq 2$, $K \geq 2$) is specified by K real-valued multivariate random variables $\mathbf{X}_1, \dots, \mathbf{X}_K$ over $[0,1]^D$. Let $\boldsymbol{\mu}_k = \mathbb{E}[\mathbf{X}_k]$ denote the expected vectorial cost of arm k where $\boldsymbol{\mu}_k = (\mu_{k,1}, \dots, \mu_{k,D})$. Furthermore, $\boldsymbol{\mu}$ denotes the matrix whose rows are the $\boldsymbol{\mu}_k$'s.

In each time step the learner can select one of the arms and obtain a sample, which is a cost vector, from the corresponding distribution. Samples are also assumed to be independent over time and across the arms, but not necessarily across the components of cost vectors. At time step t, k_t denotes the index of the arm played by the learner and $\mathbf{X}_{k_t}^{(t)} = (X_{k_t,1}^{(t)}, \dots X_{k_t,D}^{(t)})$ the resulting payoff. After playing t time steps, the empirical estimate of the expected cost $\boldsymbol{\mu}_k$ of the kth arm is:

$$\widehat{\boldsymbol{\mu}}_k^{(t)} = \frac{1}{T_k(t)} \sum_{\tau=1}^{t} \mathbf{X}_{k_\tau}^{(\tau)} \mathbf{1}(k_\tau = k) \tag{1}$$

where all operations are meant elementwise, $T_k(t)$ is the number of times the kth arm has been played (i.e., $T_k(t) = \sum_{\tau=1}^{t} \mathbf{1}(k_\tau = k)$) and $\mathbf{1}(\cdot)$ is the indicator function.

3. Multi-objective optimization

In order to complete the MO-MAB setting, we need to introduce the notion of optimality of the arms. First, we introduce the Pareto dominance relation $\preceq$ defined as follows, for any $\boldsymbol{v}, \boldsymbol{v}' \in \mathbb{R}^D$:

$$\boldsymbol{v} \preceq \boldsymbol{v}' \Leftrightarrow \forall d = 1, \dots, D, v_d \leq v'_d \ . \tag{2}$$

Let $\mathcal{O} \subseteq \mathbb{R}^D$ be a set of D-dimension vectors. The *Pareto front* of $\mathcal{O}$, denoted $\mathcal{O}^*$, is the set of vectors such that:

$$\boldsymbol{v}^* \in \mathcal{O}^* \Leftrightarrow \left(\forall \boldsymbol{v} \in \mathcal{O}, \boldsymbol{v} \preceq \boldsymbol{v}^* \Rightarrow \boldsymbol{v} = \boldsymbol{v}^*\right) \ . \tag{3}$$

In multi-objective optimization, one usually wants to compute the Pareto front, or search for a particular element of the Pareto front. In practice, it may be costly (and even infeasible depending on the size of the solution space) to determine all the solutions of the Pareto front. One may then prefer to directly aim for a particular solution in the Pareto front. This problem is formalized as a single objective optimization problem, using an *aggregation function*.

An aggregation (or scalarizing) function, which is a non-decreasing function $\phi : \mathbb{R}^D \to \mathbb{R}$, allows every vector to receive a scalar value to be optimized[2]. The initial multi-objective problem is then rewritten as follows:

$$\min \phi(\boldsymbol{v}) \quad \text{s.t.} \quad \boldsymbol{v} \in \mathcal{O} \ . \tag{4}$$

A solution to this problem yields a particular solution on the Pareto front. Note that if ϕ is not strictly increasing in every component, some care is needed to ensure that the solution of (4) is on the Pareto front.

Different aggregation function can be used depending on the problem at hand, such as sum, weighted sum, $\min$, $\max$, (augmented) weighted Chebyshev norm (Steuer & Choo, 1983), Ordered Weighted Averages (OWA) (Yager, 1988) or Ordered Weighted Regret (OWR) (Ogryczak et al., 2011) and its weighted version (Ogryczak et al., 2013). In this study, we focus on the Generalized Gini Index (GGI) (Weymark, 1981), a special case of OWA.

[1] Our setup is motivated by a practical application where feedback is more naturally formulated in terms of cost. However the stochastic bandit problem is generally based on rewards, which can be easily turned into costs by using the transformation $x \mapsto 1 - x$ assuming that the rewards are from $[0, 1]$.

[2] A multivariate function $f : \mathbb{R}^D \to \mathbb{R}$ is said to be monotone (non-decreasing) if for all fixed $\mathbf{x}, \mathbf{x}' \in \mathbb{R}^D$ such that $\mathbf{x} \preceq \mathbf{x}'$ implies that $f(\mathbf{x}) \leq f(\mathbf{x}')$.

4. Generalized Gini Index

For a given $n \in \mathbb{N}$, $[n]$ denotes the set $\{1, 2, \ldots, n\}$. The Generalized Gini Index (GGI) (Weymark, 1981) is defined as follows for a cost vector $\mathbf{x} = (x_1, \ldots, x_D) \in \mathbb{R}^D$:

$$G_{\mathbf{w}}(\mathbf{x}) = \sum_{d=1}^{D} w_d x_{\sigma(d)} = \mathbf{w}^\intercal \mathbf{x}_\sigma$$

where $\sigma \in \mathbb{S}_D$, which depends on $\mathbf{x}$, is the permutation that sorts the components of $\mathbf{x}$ in a decreasing order, $\mathbf{x}_\sigma = (x_{\sigma(1)}, \cdots, x_{\sigma(D)})$ is the sorted vector and weights w_i's are assumed to be non-increasing, i.e., $w_1 \geq w_2 \geq \ldots \geq w_D$. Given this assumption, $G_{\mathbf{w}}(\mathbf{x}) = \max_{\pi \in \mathbb{S}_D} \mathbf{w}^\intercal \mathbf{x}_\pi = \max_{\pi \in \mathbb{S}_D} \mathbf{w}_\pi^\intercal \mathbf{x}$ and is therefore a *piecewise-linear convex* function. Figure 1 illustrates GGI on a bi-objective optimization task.

To better understand GGI, we introduce its formulation in terms of Lorenz vectors. The Lorenz vector of $\mathbf{x}$ is the vector $\mathbf{L}(\mathbf{x}) = (L_1(\mathbf{x}), \ldots, L_D(\mathbf{x}))$ where $L_d(\mathbf{x})$ is the sum of the d smallest components of $\mathbf{x}$. Then, GGI can be rewritten as follows:

$$G_{\mathbf{w}}(\mathbf{x}) = \sum_{d=1}^{D} w'_d L_d(\mathbf{x}) \tag{5}$$

where $\mathbf{w}' = (w'_1, \ldots, w'_D)$ is the vector defined by $\forall d \in [D], w'_d = w_d - w_{d+1}$ with $w_{D+1} = 0$. Note that all the components of $\mathbf{w}'$ are nonnegative as we assume that those of $\mathbf{w}$ are non-increasing.

GGI[3] was originally introduced for quantifying the inequality of income distribution in economics. It is also known in statistics (Buczolich & Székely, 1989) as a special case of Weighted Average Ordered Sample statistics, which does not require that weights be non-increasing and is therefore not necessarily convex. GGI has been characterized by Weymark (1981). It encodes both efficiency as it is monotone with Pareto dominance and fairness as it is non-increasing with Pigou-Dalton transfers (1912; 1920); they are two principles formulating natural requirements, which is an important reason why GGI became a well-established measure of balancedness. Informally, a Pigou-Dalton transfer amounts to increasing a lower-valued objective while decreasing another higher-valued objective by the same quantity such that the order between the two objectives is not reversed. The effect of such a transfer is to balance a cost vector. Formally, GGI satisfies the following fairness property: $\forall \mathbf{x}$ such that $x_i < x_j$,

$$\forall \epsilon \in (0, x_j - x_i), G_{\mathbf{w}}(\mathbf{x} + \epsilon \boldsymbol{e}_i - \epsilon \boldsymbol{e}_j) \leq G_{\mathbf{w}}(\mathbf{x})$$

where $\boldsymbol{e}_i$ and $\boldsymbol{e}_j$ are two vectors of the canonical basis. As a consequence, among vectors of equal sum, the best cost vector (w.r.t. GGI) is the one with *equal values in all objectives* if feasible.

[3]Note that in this paper GGI is expressed in terms of costs and therefore lower GGI values are preferred.

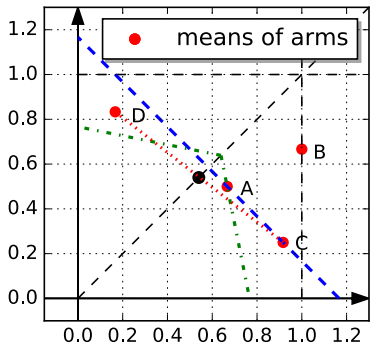

Figure 1. Point A is always preferred (w.r.t. GGI) to B due to Pareto dominance; A is always preferred to C due to the Pigou-Dalton transfer principle; depending on the weights of GGI, A may be preferred to D or the other way around; with $\mathbf{w} = (1, 1/2)$, A is preferred to D (points equivalent to A are on the dashed green line). The optimal mixed solution is the black dot.

The original Gini index can be recovered as a special case of GGI by setting the weights as follows:

$$\forall d \in [D], w_d = (2(D - d) + 1)/D^2. \tag{6}$$

This yields a nice graphical interpretation. For a fixed vector $\mathbf{x}$, the distribution of costs can be represented as a curve connecting the points $(0, 0)$, $(1/D, L_1(\mathbf{x}))$, $(2/D, L_2(\mathbf{x}))$, $\ldots$, $(1, L_D(\mathbf{x}))$. An ideally fair distribution with an identical total cost is given by the straight line connecting $(0, 0)$, $(1/D, L_D(\mathbf{x})/D)$, $(2/D, 2L_D(\mathbf{x})/D)$, $\ldots$, $(1, L_D(\mathbf{x}))$, which equally distributes the total cost over all components. Then $1 - G_{\mathbf{w}}(\mathbf{x})/\bar{x}$ with weights (6) and $\bar{x} = \sum x_i/D$ is equal to twice the area between the two curves.

From now on, to simplify the presentation, we focus on GGI with strictly decreasing weights in $[0, 1]^D$, i.e., $d < d'$ implies $w_d > w_{d'}$. This means that GGI is strictly decreasing with Pigou-Dalton transfers and all the components of $\mathbf{w}'$ are positive. Based on formulation (5), Ogryczak & Sliwinski (2003) showed that the GGI value of a vector $\mathbf{x}$ can be obtained by solving a linear program. We shall recall their results and define the linear program-based formulation of GGI.

Proposition 1. *The* GGI *score* $G_{\mathbf{w}}(\mathbf{x})$ *of vector* $\mathbf{x}$ *is the optimal value of the following linear program*

$$\begin{array}{lllll} \textit{minimize} & \displaystyle\sum_{d=1}^{D} w'_d \left(d r_d + \sum_{j=1}^{D} b_{j,d} \right) & & \\ \textit{subject to} & r_d + b_{j,d} & \geq x_j & \forall j, d \in [D] \\ & b_{j,d} & \geq 0 & \forall j, d \in [D] \end{array}$$

5. Optimal policy

In the single objective case, arms are compared in terms of their means, which induce a total ordering over arms. In the multi-objective setting, we use the GGI criterion to compare arms. One can compute the GGI score of each arm k as $G_{\mathbf{w}}(\boldsymbol{\mu}_k)$ if its vectorial mean $\boldsymbol{\mu}_k$ is known. Then an optimal arm k^* minimizes the GGI score, i.e.,

$$k^* \in \operatorname*{argmin}_{k\in[K]} G_{\mathbf{w}}(\boldsymbol{\mu}_k) \ .$$

However, in this work, we also consider *mixed* strategies, which can be defined as $\mathcal{A} = \{\boldsymbol{\alpha} \in \mathbb{R}^K | \sum_{k=1}^K \alpha_k = 1 \wedge 0 \preceq \boldsymbol{\alpha}\}$, because they may allow to reach lower GGI values than any fixed arm (see Figure 1). A policy parameterized by $\boldsymbol{\alpha}$ chooses arm k with probability α_k. An optimal mixed policy can then be obtained as follows:

$$\boldsymbol{\alpha}^* \in \operatorname*{argmin}_{\boldsymbol{\alpha}\in A} G_{\mathbf{w}}\left(\sum_{k=1}^K \alpha_k \boldsymbol{\mu}_k\right) \ . \tag{7}$$

In general, $G_{\mathbf{w}}\left(\sum_{k=1}^K \alpha_k^* \boldsymbol{\mu}_k\right) \leq G_{\mathbf{w}}(\boldsymbol{\mu}_{k^*})$, therefore using mixed strategies is justified in our setting. Based on Proposition 1, if the arms' means were known, $\boldsymbol{\alpha}^*$ could be computed by solving the following linear program:

$$\begin{aligned} \text{minimize} \quad & \sum_{d=1}^D w'_d \left(d r_d + \sum_{j=1}^D b_{j,d} \right) \\ \text{subject to} \quad & r_d + b_{j,d} \geq \sum_{k=1}^K \alpha_k \mu_{k,j} \quad \forall j,d \in [D] \\ & \textstyle\sum_{k=1}^K \alpha_k = 1 \qquad \boldsymbol{\alpha} \succeq \mathbf{0} \\ & b_{j,d} \geq 0 \qquad\qquad\qquad \forall j,d \in [D] \end{aligned} \tag{8}$$

6. Regret

After playing T rounds, the average cost can be written as

$$\bar{\mathbf{X}}^{(T)} = \frac{1}{T}\sum_{t=1}^T \mathbf{X}_{k_t}^{(t)} \ .$$

Our goal is to minimize the GGI index of this term. Accordingly we expect the learner to collect costs so as their average in terms of GGI, that is, $G_{\mathbf{w}}\left(\bar{\mathbf{X}}^{(T)}\right)$ should be as small as possible. As shown in the previous section, for a given bandit instance with arm means $\boldsymbol{\mu} = (\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K)$, the optimal policy $\boldsymbol{\alpha}^*$ achieves $G_{\mathbf{w}}\left(\sum_{k=1}^K \alpha_k^* \boldsymbol{\mu}_k\right) = G_{\mathbf{w}}\left(\boldsymbol{\mu}\boldsymbol{\alpha}^*\right)$ if the randomness of the costs are not taken into account. We consider the performance of the optimal policy as a reference value, and define the regret of the learner as the difference of the GGI of its average cost and the GGI of the optimal policy:

$$R^{(T)} = G_{\mathbf{w}}\left(\bar{\mathbf{X}}^{(T)}\right) - G_{\mathbf{w}}\left(\boldsymbol{\mu}\boldsymbol{\alpha}^*\right) \ . \tag{9}$$

Note that GGI is a continuous function, therefore if the learner follows a policy $\boldsymbol{\alpha}^{(T)}$ that is "approaching" $\boldsymbol{\alpha}^*$ as $T \to \infty$, then the regret is vanishing.

We shall also investigate a slightly different regret notion called *pseudo-regret*:

$$\overline{R}^{(T)} = G_{\mathbf{w}}\left(\boldsymbol{\mu}\bar{\boldsymbol{\alpha}}^{(T)}\right) - G_{\mathbf{w}}\left(\boldsymbol{\mu}\boldsymbol{\alpha}^*\right) \tag{10}$$

where $\bar{\boldsymbol{\alpha}}^{(T)} = \frac{1}{T}\sum_{t=1}^T \boldsymbol{\alpha}^{(t)}$. We will show that the difference between the regret and pseudo-regret of our algorithm is $\tilde{O}(T^{-1/2})$ with high probability, thus having a high probability regret bound $\tilde{O}(T^{-1/2})$ for one of them implies a regret bound $\tilde{O}(T^{-1/2})$ for the other one.

Remark 1. *The single objective stochastic multi-armed bandit problem (Auer et al., 2002a; Bubeck & Cesa-Bianchi, 2012) can be naturally accommodated into our setup with $D = 1$ and $w_1 = 1$. In this case, $\boldsymbol{\alpha}^*$ implements the pure strategy that always pulls the optimal arm with the highest mean denoted by μ_{k^*}. Thus $G_{\mathbf{w}}\left(\boldsymbol{\mu}\boldsymbol{\alpha}^*\right) = \mu_{k^*}$ in this case. Assuming that the learner plays only with pure strategies, the pseudo-regret defined in (10) can be written as:*

$$\overline{R}^{(T)} = \frac{1}{T}\sum_{t=1}^T \left(\mu_{k_t} - T\mu_{k^*}\right) = \frac{1}{T}\sum_{k=1}^K T_k(T)\left(\mu_k - \mu_{k^*}\right)$$

which coincides with the single objective pseudo-regret (see for example (Auer et al., 2002a)), apart from the fact that we work with costs instead of rewards. Therefore our notion of multi-objective pseudo-regret can be seen as a generalization of the single objective pseudo-regret.

Remark 2. *Single-objective bandit algorithm can be applied in our multi-objective setting by transforming the multivariate payoffs $\mathbf{X}_{k_t}^{(t)}$ into a single real value $G_{\mathbf{w}}(\mathbf{X}_{k_t}^{(t)})$ in every time step t. However, in general, this approach fails to optimize GGI as formulated in (7) due to GGI's non-linearity, even if the optimal policy is a pure strategy. Moreover, applying a multi-objective bandit algorithm such as MO-UCB (Drugan & Nowé, 2013) would be inefficient as they were developed to find all Pareto-optimal arms and then to sample them uniformly at random. This approach may be reasonable to apply only when $\alpha_k^* = 1/\#\mathcal{K}$ where $\mathcal{K} = \{k \in [K] : \alpha_k^* > 0\}$ contains all the Pareto optimal arms, which is clearly not the case for every MO-MAB instance.*

7. Learning algorithm based on OCO

In this section we propose an online learning algorithm called MO-OGDE, to optimize the regret defined in the previous section. Our method exploits the convexity of the GGI operator and formalizes the policy search problem as an online convex optimization problem, which is solved by Online Gradient Descent (OGD) (Zinkevich, 2003) algorithm with projection to a gradually expanding truncated probability simplex.

Algorithm 1 MO-OGDE(δ)

1: Pull each arm once
2: Set $\boldsymbol{\alpha}^{(K+1)} = (1/K, \cdots, 1/K)$
3: **for** rounds $t = K+1, K+2, \ldots$ **do**
4: Choose an arm k_t according to $\boldsymbol{\alpha}^{(t)}$
5: Observe the sample $\mathbf{X}_{k_t}^{(t)}$ and compute $f^{(t)}$
6: Set $\eta_t = \frac{\sqrt{2}}{(1-1/\sqrt{K})}\sqrt{\frac{\ln(2/\delta)}{t}}$
7: $\boldsymbol{\alpha}^{(t+1)} = \mathrm{OGDEstep}(\boldsymbol{\alpha}^{(t)}, \eta_t, \nabla f^{(t)})$
return $\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{\alpha}^{(t)}$

Then we shall provide a regret analysis of our method. Due to space limitation, the proofs are deferred to the appendix.

7.1. MO-OGDE

Our objective function to be minimized can be viewed as a function of $\boldsymbol{\alpha}$, i.e., $f(\boldsymbol{\alpha}) = G_{\mathbf{w}}(\boldsymbol{\mu}\boldsymbol{\alpha})$ where the matrix $\boldsymbol{\mu} = (\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K)$ contains the means of the arm distributions as its columns. Note that the convexity of GGI implies the convexity of $f(\boldsymbol{\alpha})$. Since we play with mixed strategies, the domain of our optimization problem is the K-dimensional probability simplex $\Delta_K = \{\boldsymbol{\alpha} \in \mathbb{R}^K | \sum_{k=1}^{K}\alpha_k = 1 \wedge 0 \preceq \boldsymbol{\alpha}\}$, which is a convex set. Then the gradient of $f(\boldsymbol{\alpha})$ with respect to α_k can be computed as $\frac{\partial f(\boldsymbol{\alpha})}{\partial \alpha_k} = \sum_{d=1}^{D} w_d \mu_{k,\pi(d)}$ where π is the permutation that sorts the components of $\boldsymbol{\mu}\boldsymbol{\alpha}$ in a decreasing order. The means $\boldsymbol{\mu}_k$'s are not known but they can be estimated based on the costs observed so far as given in (1). The objective function based on the empirical mean estimates is denoted by $f^{(t)}(\boldsymbol{\alpha}) = G_{\mathbf{w}}(\widehat{\boldsymbol{\mu}}^{(t)}\boldsymbol{\alpha})$ where $\widehat{\boldsymbol{\mu}}^{(t)} = (\widehat{\boldsymbol{\mu}}_1^{(t)}, \ldots, \widehat{\boldsymbol{\mu}}_K^{(t)})$ contains the empirical estimates for $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K$ in time step t as its columns.

Our Multi-Objective Online Gradient Descent algorithm with Exploration is defined in Algorithm 1, which we shall refer to as MO-OGDE. Our algorithm is based on the well-known Online Gradient Descent (Zinkevich, 2003; Hazan, 2016) that carries out the gradient step and the projection back onto the domain in each round. The MO-OGDE algorithm first pulls each arm at once as an initialization step. Then in each iteration, it chooses each arm k with probability $\alpha_k^{(t)}$, and it computes $f^{(t)}$ based on the empirical mean estimates. Next, it carries out the gradient step based on $\nabla f^{(t)}(\boldsymbol{\alpha}^{(t)})$ with a step size η_t as defined in line 6, and computes the projection $\Pi_{\Delta_K^\beta}$ onto the nearest point of the convex set:

$$\Delta_K^\beta = \left\{\boldsymbol{\alpha} \in \mathbb{R}^K | \sum_{k=1}^{K}\alpha_k = 1 \wedge \beta/K \preceq \boldsymbol{\alpha}\right\}$$

with $\beta = \eta_t$. The gradient step and projection are carried out using

$$\mathrm{OGDEstep}(\boldsymbol{\alpha}, \eta, g) = \Pi_{\Delta_K^\eta}(\boldsymbol{\alpha} - \eta g(\boldsymbol{\alpha})) \qquad (11)$$

The key ingredient of MO-OGDE is the projection step onto the truncated probability simplex $\Delta_K^{\eta_t}$ which ensures that $\alpha_k^{(t)} > \eta_t/K$ for every $k \in [K]$ and $t \in [T]$. This forced exploration is indispensable in our setup, since the objective function $f(\boldsymbol{\alpha})$ depends on the means of the arm distributions, which are not known. To control the difference between $f(\boldsymbol{\alpha})$ and $f^{(t)}(\boldsymbol{\alpha})$, we need "good" estimates for $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K$, which are obtained via forced exploration. That is why, the original OGD (Hazan, 2016) algorithm, in general, fails to optimize GGI in our setting. Our analysis, presented in the next section, focuses on the interplay between the forced exploration and the bias of the loss functions $f^{(1)}(\boldsymbol{\alpha}), \ldots, f^{(T)}(\boldsymbol{\alpha})$.

7.2. Regret analysis

The technical difficulty in optimizing GGI in an online fashion is that, in general, $f^{(t)}(\boldsymbol{\alpha}) - f(\boldsymbol{\alpha})$ $(= G_{\mathbf{w}}(\widehat{\boldsymbol{\mu}}^{(t)}\boldsymbol{\alpha}) - G_{\mathbf{w}}(\boldsymbol{\mu}\boldsymbol{\alpha}))$ is of order $\min_k(T_k(t))^{-1/2}$, which, unless all the arms are sampled a linear number of times, incurs a regret of the same order of magnitude, which is typically too large. Nevertheless, an optimal policy $\boldsymbol{\alpha}^*$ determines a convex combination of several arms in the form of $\boldsymbol{\mu}\boldsymbol{\alpha}^*$, which is the optimal cost vector in terms of GGI given the arm distribution at hand. Let us denote $\mathcal{K} = \{k \in [K] : \alpha_k^* > 0\}$. Note that $\#\mathcal{K} \leq D$. Moreover, arms in $[K] \setminus \mathcal{K}$ with an individual GGI lower than arms in $\mathcal{K}$ do not necessarily participate in the optimal combination. Obviously, an algorithm that achieves a $O(T^{-1/2})$ needs to pull the arms in $\#\mathcal{K}$ linear time, and at the same time, estimate the arms in $[K] \setminus \mathcal{K}$ with a certain precision.

The main idea in the approach proposed in this paper is, despite the above remarks, to apply some online convex optimization algorithm on the current estimate $f^{(t)}(\boldsymbol{\alpha}) = G_{\mathbf{w}}(\widehat{\boldsymbol{\mu}}^{(t)}\boldsymbol{\alpha})$ of the objective function $f(\boldsymbol{\alpha}) = G_{\mathbf{w}}(\boldsymbol{\mu}\boldsymbol{\alpha})$, use forced exploration of order $T^{1/2}$, and finally show that the estimate $f^{(t)}$ of the objective function has error $\tilde{O}(T^{-1/2})$ *along the trajectory* generated by the online convex optimization algorithm. In particular, we show that $f(\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{\alpha}^{(t)}) \leq \frac{1}{T}\sum_{t=1}^{T} f^{(t)}(\boldsymbol{\alpha}^{(t)}) + \tilde{O}(T^{-1/2})$. The intuitive reason for this is that $\|\frac{1}{T}\sum_{t=1}^{T}\widehat{\boldsymbol{\mu}}^{(t)}\boldsymbol{\alpha}^{(t)} - \frac{1}{T}\sum_{t=1}^{T}\boldsymbol{\mu}\boldsymbol{\alpha}^{(t)}\| = \tilde{O}(T^{-1/2})$, which is based on the following observation: an arm in $\mathcal{K}$ is either pulled often, thus its mean estimate is then accurate enough, or an arm in $[K] \setminus \mathcal{K}$ is pulled only a few times, nevertheless $\sum_{t=1}^{T}\alpha_k^{(t)}$ is then small enough to make the poor accuracy of its mean estimate insignificant. Below we make this argument formal by proving the following theorem:

Theorem 1. *With probability at least* $1-\delta$*:*

$$f\Big(\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{\alpha}^{(t)}\Big) - f(\boldsymbol{\alpha}^*) \leq 2L\sqrt{\frac{6D\ln^3(8DKT^2/\delta)}{T}}\ ,$$

for any big enough T*, where* L *is the Lipschitz constant of*

$G_{\mathbf{w}}(\mathbf{x})$.

Its proof follows four subsequent steps presented next.

Step 1 As a point of departure, we analyze OGDEstep in (11). In particular, we compute the regret that is commonly used in OCO setup for $f^{(t)}(\boldsymbol{\alpha}^{(t)})$.

Lemma 1. *Define* $\{\boldsymbol{\alpha}^{(t)}\}_{t=1,\dots,T}$ *as:*

$$\boldsymbol{\alpha}^{(1)} = (1/K, \dots, 1/K)$$
$$\boldsymbol{\alpha}^{(t+1)} = \text{OGDEstep}(\boldsymbol{\alpha}^{(t)}, \eta_t, \nabla f^{(t)})$$

with $\eta_1, \dots, \eta_T \in [0,1]$. *Then the following upper bound is guaranteed for all* $T \geq 1$ *and for any* $\boldsymbol{\alpha} \in \Delta_K$*:*

$$\sum_{t=1}^{T} f^{(t)}(\boldsymbol{\alpha}^{(t)}) - \sum_{t=1}^{T} f^{(t)}(\boldsymbol{\alpha}) \leq \frac{1}{\eta_T} + \frac{G^2+1}{2}\sum_{t=1}^{T}\eta_t$$

where $\sup_{\boldsymbol{\alpha}\in\Delta_K} \|\nabla f^{(t)}(\boldsymbol{\alpha})\| < G \leq \sqrt{K}D$ *for all* $t \in [T]$.

The proof of Lemma 1 is presented in Appendix A. If the projection is carried out onto Δ_K according to the OGE algorithm instead of the truncated probability $\Delta_K^{\eta_t}$, the regret bound in Lemma 1 would be improved only by a constant factor (see Theorem 3.1 in (Hazan, 2016)). As a side remark, note that Lemma 1 holds for arbitrary convex function since only the convexity of $f^{(t)}(\boldsymbol{\alpha})$ is used in the proof.

Step 2 Next, we show that $f^{(t)}(\boldsymbol{\alpha})$ converges to $f(\boldsymbol{\alpha})$ as fast as $\tilde{O}(T^{-1/2})$ along the trajectory $\{\boldsymbol{\alpha}^{(t)}\}_{t=1,\dots,T}$ generated by MO-OGDE.

Proposition 2. *With probability at least* $1 - 2(DT+1)K\delta$,

$$\left| G_{\mathbf{w}}\left(\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{\mu}\boldsymbol{\alpha}^{(t)}\right) - G_{\mathbf{w}}\left(\frac{1}{T}\sum_{t=1}^{T}\widehat{\boldsymbol{\mu}}^{(t)}\boldsymbol{\alpha}^{(t)}\right)\right| \leq L\sqrt{\frac{6D(1+\ln^2 T)\ln(2/\delta)}{T}}.$$

The proof of Proposition 2 is deferred to Appendix B. Proposition 2, combined with the fact that

$$G_{\mathbf{w}}\left(\frac{1}{T}\sum_{t=1}^{T}\widehat{\boldsymbol{\mu}}^{(t)}\boldsymbol{\alpha}^{(t)}\right) \leq \frac{1}{T}\sum_{t=1}^{T} G_{\mathbf{w}}(\widehat{\boldsymbol{\mu}}^{(t)}\boldsymbol{\alpha}^{(t)}) = \frac{1}{T}\sum_{t=1}^{T} f^{(t)}(\boldsymbol{\alpha}^{(t)})$$

where we used the convexity of GGI, and $G_{\mathbf{w}}\left(\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{\mu}\boldsymbol{\alpha}^{(t)}\right) = f\left(\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{\alpha}^{(t)}\right) = f\left(\bar{\boldsymbol{\alpha}}^{(t)}\right)$ implies the following result.

Corollary 1. *With probability at least* $1 - 2(DT+1)K\delta$,

$$f\left(\bar{\boldsymbol{\alpha}}^{(t)}\right) \leq \frac{1}{T}\sum_{t=1}^{T} f^{(t)}\left(\boldsymbol{\alpha}^{(t)}\right) + L\sqrt{\frac{6D(1+\ln^2 T)\ln\frac{2}{\delta}}{T}}.$$

Step 3 Next, we provide a regret bound for the pseudo-regret of MO-OGDE by using Lemma 1 and Corollary 1. To this end, we introduce some further notations. First of all, let

$$\Delta_K^* = \operatorname*{argmin}_{\boldsymbol{\alpha}\in\Delta_K} f(\boldsymbol{\alpha})$$

denote the set of all the minimum points of f over Δ_K. As we show later in Lemma 3, Δ_K^* is a convex polytope, and thus the set $\text{ext}(\Delta_K^*)$ of its extreme points is finite.

Proposition 3. *With probability at least* $1 - 4DT^2K\delta$*:*

$$f\left(\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{\alpha}^{(t)}\right) - f(\boldsymbol{\alpha}^*) \leq L\sqrt{\frac{6D(1+\ln^2 T)\ln(2/\delta)}{T}} + \frac{1}{T\eta_T} + \frac{KD^2+1}{T}\sum_{t=1}^{T}\eta_t + \frac{LK}{T}\sum_{t=1}^{T}\sqrt{\frac{D\ln(2/\delta)}{2\chi_1(t)}}$$

where $\chi_1(t) = \mathbf{1}(t \leq \tau_1) + \mathbf{1}(t > \tau_1)(ta_0/(2|\text{ext}(\Delta_K^*)|))$ *and* $\tau_1 = \left\lceil (2|\text{ext}(\Delta_K^*)|)\left[2 + \frac{10\sqrt{3}LKD^2}{g^*}\right]\sqrt{2\ln\frac{2}{\delta}}\right\rceil^4$ *with* $a_0 = \min_{\boldsymbol{\alpha}\in\text{ext}(\Delta_K^*)}\min_{k:\alpha_k>0}\alpha_k$ *and* $g^* = \inf_{\boldsymbol{\alpha}\in\Delta_K\setminus\Delta_K^*}\max_{\boldsymbol{\alpha}^*\in\Delta_K^*}\frac{f(\boldsymbol{\alpha})-f(\boldsymbol{\alpha}^*)}{\|\boldsymbol{\alpha}-\boldsymbol{\alpha}^*\|}$.

The proof is deferred to Appendix C. In the proof first we decompose the regret into various terms which can be upper bounded based on Lemma 1 and Corollary 1. This implies that $\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{\alpha}_k^{(t)}$ will get arbitrarily close to some $\boldsymbol{\alpha}_k^* \in \Delta_K^*$ if T is big enough because, as we show later in Lemma 3, g^* is strictly positive (see Appendix D). As a consequence, for a big enough $T > \tau_1$, the difference between the $f^{(t)}(\boldsymbol{\alpha}^*)$ and $f(\boldsymbol{\alpha}^*)$ is vanishing as fast as $O(T^{-1/2})$ which is crucial for a regret bound of order $T^{-1/2}$.

Step 4 Finally, Theorem 1 yields from Proposition 3 by simplifying the right hand side. The calculation is presented in Appendix E.

7.3. Regret vs. pseudo-regret

Next, we upper-bound the difference of regret defined in (9) and the pseudo-regret defined in (10). To that aim, we first upper-bound the difference of $T_k(t)$ and $\sum_{\tau=1}^{t}\alpha_k^{(\tau)}$.

Claim 1. *For any* $t = 1, 2, \dots$ *and any* $k = 1, \dots, K$ *it holds that* $\mathbb{P}\left[\left|T_k(t) - \sum_{\tau=1}^{t}\alpha_k^{(\tau)}\right| \geq \sqrt{2t\ln(2/\delta)}\right] \leq \delta$.

Proof : As $\mathbb{P}[k_t = k] = \alpha_k$, it holds that $T_k(t) - \sum_{\tau=1}^{t}\alpha_k^{(\tau)} = \sum_{\tau=1}^{t}[\mathbf{1}(k_\tau = k) - \alpha_k^{(\tau)}]$ is a martingale. Besides, $|\mathbf{1}(k_\tau = k) - \alpha_k| \leq 1$ by construction. The claim then follows by Azuma's inequality. ■

Based on Claim 1 and Prop. 2, we upper-bound the difference between the pseudo-regret and regret of MO-OGDE.

Corollary 2. *With probability at least* $1-\delta$

$$|R^{(T)} - \bar{R}^{(T)}| \leq L\sqrt{\frac{12D\ln(4(DT+1)/\delta)}{T}}$$

The proof of Corollary 2 is deferred to Appendix F. According to Corollary 2, the difference between the regret and pseudo regret is $\tilde{O}(T^{-1/2})$ with high probability, hence Theorem 1 implies a $\tilde{O}(T^{-1/2})$ bound for the regret of MO-OGDE.

8. Experiments

To test our algorithm, we carried out two sets of experiments. In the first we generated synthetic data from multi-objective bandit instances with known parameters. In this way, we could compute the pseudo-regret (10) and, thus investigate the empirical performance of the algorithms. In the second set of experiments, we run our algorithm on a complex multi-objective online optimization problem, namely an electric battery control problem. Before presenting those experiments, we introduce another algorithm that will serve as a baseline.

8.1. A baseline method

In the previous section we introduced a gradient-based approach that uses the mean estimates to approximate the gradient of the objective function. Nevertheless, using the mean estimates, the optimal policy can be directly approximated by solving the the linear program given in (8). We use the same exploration as MO-OGDE, see line 6 of Algorithm 1. More concretely, the learner solves the following linear program in each time step t:

$$\begin{aligned}
\text{minimize} \quad & \sum_{d=1}^{D} w'_d \left(d r_d + \sum_{j=1}^{D} b_{j,d} \right) \\
\text{subject to} \quad & r_d + b_{j,d} \geq \sum_{k=1}^{K} \alpha_k \widehat{\mu}_{k,j}^{(t)} \quad \forall j, d \in [D] \\
& \boldsymbol{\alpha}^T \mathbf{1} = 1 \\
& \boldsymbol{\alpha} \geq \eta_t / K \\
& b_{j,d} \geq 0 \qquad \forall j, d \in [D]
\end{aligned}$$

Note that the solution of the learner program above regarding $\boldsymbol{\alpha}$ is in $\Delta_K^{\eta_t}$. We refer to this algorithm as MO-LP. Note that this approach is computationally expensive, since a linear program needs to be solved at each time step. But the policy of each step is always optimal restricted to the truncated simplex $\Delta_K^{\eta_t}$ with respect to the mean estimates, unlike the gradient descent method.

8.2. Synthetic Experiments

We generated random multi-objective bandit instances for which each component of the multivariate cost distributions obeys Bernoulli distributions with various parameters. The parameters of each Bernoulli distributions are drawn uniformly at random from $[0, 1]$ independently from each other. The number of arms K was set to $\{5, 20\}$ and the dimension of the cost distribution was taken from $D \in \{5, 10\}$. The weight vector $\mathbf{w}$ of GGI was set to $w_d = 1/2^{d-1}$. Since the parameters of the bandit instance are known, the regret defined in Section 6 can be computed. We ran the MO-OGDE and MO-LP algorithms with 100 repetitions. The multi-objective bandit instance were regenerated after each run. The regrets of the two algorithms, which are averaged out over the repetitions, are plotted in Figure 2 along with the error bars. The results reveal some general trends. First, the average regrets of both algorithms converge to zero. Second the MO-LP algorithm outperforms the gradient descent algorithm for small number of round, typically $T < 5000$ on the more complex bandit instances ($K = 20$). This fact might be explained by the fact that the MO-LP solves a linear program for estimating α^* whereas the MO-OGDE minimizes the same objective but using a gradient descent approach with projection, which might achieve slower convergence in terms of regret, nevertheless its computational time is significantly decreased compared to the baseline method.

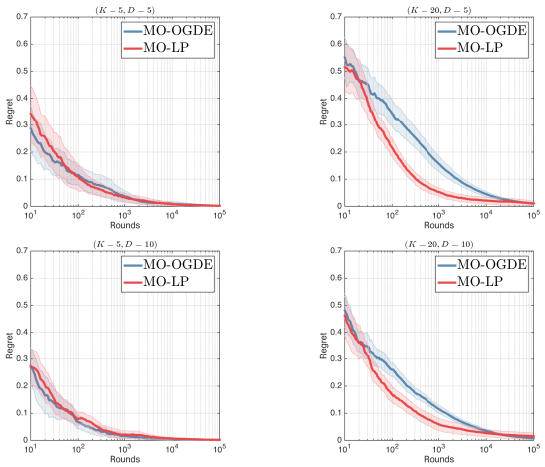

Figure 2. The regret of the MO-LP and MO-OGDE. The regret and its error is computed based on 100 repetitions and plotted in terms of the number of rounds. The dimension of the arm distributions was set to $D \in \{5, 10\}$, which is indicated in the title of the panels.

8.3. Battery control task

We also tried our algorithms on a more realistic domain: the cell balancing problem. As the performance profile of battery cells, subcomponents of an electric battery, may vary due to small physical and manufacturing differences, efficient balancing of those cells is needed for better performance and longer battery life. We model this problem as a MO-MAB where the arms are different cell control strategies and the

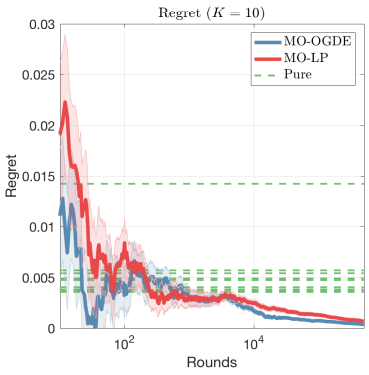

Figure 3. The regret of the MO-OGDE and MO-LP on the battery control task. The regret is averaged over 100 repetitions and plotted in terms of the number of rounds. The dimension of the arm distributions was $D = 12$.

goal is to balance several objectives: state-of-charge (SOC), temperature and aging. More concretely, the learner loops over the following two steps: (1) she chooses a control strategy for a short duration and (2) she observes its effect on the objectives (due to stochastic electric consumption). For the technical details of this experiments see Appendix G.

We tackled this problem as a GGI optimization problem. The results (averaged over 100 runs) are presented in Figure 3 where we evaluated MO-OGDE vs. MO-LP. The dashed green lines represent the regrets of playing fixed deterministic arms. Although MO-OGDE and MO-LP both learn to play a mixed policy that is greatly better than any individual arm, MO-OGDE is computationally much more efficient.

9. Related work

The single-objective MAB problem has been intensively studied especially in recent years (Bubeck & Cesa-Bianchi, 2012), nevertheless there is only a very limited number of work concerning the multi-objective setting. To the best of our best knowledge, Drugan & Nowé (2013) considered first the multi-objective multi-armed problem in a regret optimization framework with a stochastic assumption. Their work consists of extending the UCB algorithm (Auer et al., 2002a) so as to be able to handle multi-dimensional feedback vectors with the goal of determining all arms on the Pareto front.

Azar et al. (2014) investigated a sequential decision making problem with vectorial feedback. In their setup the agent is allowed to choose from a finite set of actions and then it observes the vectorial feedback for each action, thus it is a full information setup unlike our setup. Moreover, the feedback is non-stochastic in their setup, as it is chosen by an adversary. They propose an algorithm that can handle a general class of aggregation functions, such as the set of bounded domain, continuous, Lipschitz and quasi-concave functions.

In the online convex optimization setup with multiple objectives (Mahdavi et al., 2013), the learner's forecast $\mathbf{x}^{(t)}$ is evaluated in terms of multiple convex loss functions $f_0^{(t)}(\mathbf{x}), f_1^{(t)}(\mathbf{x}), \ldots, f_K^{(t)}(\mathbf{x})$ in each time step t. The goal of the learner is then to minimize $\sum_{\tau=1}^{t} f_0^{(\tau)}(\mathbf{x}^{(\tau)})$ while keeping the other objectives below some predefined threshold, i.e. $\frac{1}{t}\sum_{\tau=1}^{t} f_i^{(\tau)}(\mathbf{x}^{(\tau)}) \leq \gamma_i$ for all $i \in [K]$. Note that, with linear loss functions, the multiple-objective convex optimization setup boils down to linear optimization with stochastic constraints, and thus it can be applied to solve the linear program given in Section 5 whose solution is the optimal policy in our setup. For doing this, however, each linear stochastic constraint needs to be observed, whereas we only assume bandit information.

In the approachability problem (Mannor et al., 2014; 2009; Abernethy et al., 2011), there are two players, say A and B. Players A and B choose actions from the compact convex sets $\mathcal{X} \subset \mathbb{R}^K$ and $\mathcal{Y} \subset \mathbb{R}^D$, respectively. The feedback to the players is computed as a function $u : \mathcal{X} \times \mathcal{Y} \mapsto \mathbb{R}^p$. A given convex set $S \subset \mathbb{R}^p$ is *known to both players*. Player A wants to land inside with the cumulative payoffs, i.e., player A's goal is to minimize $\text{dist}(\frac{1}{T}\sum_{t=1}^{T} u(\mathbf{x}^{(t)}, \mathbf{y}^{(t)}), S)$ where $\mathbf{x}^{(t)}$ and $\mathbf{y}^{(t)}$ are the actions chosen by player A and B respectively, and $\text{dist}(\mathbf{s}, S) = \inf_{\mathbf{s}' \in S} \|\mathbf{s}' - \mathbf{s}\|$, whereas player B, called adversary, wants to prevent player A to land in set S. In our setup, Player B who generates the cost vectors, is assumed to be stochastic. The set S consists of a single value which is $\boldsymbol{\mu}\boldsymbol{\alpha}^*$, and u corresponds to $\boldsymbol{\mu}Ibalpha$, thus $p = D$. What makes our setup essentially different from approachability is that, S is not known to any player. That is why Player A, i.e. the learner, needs to explore the action space which is achieved by forced exploration.

10. Conclusion and future work

We introduced a new problem in the context of multi-objective multi-armed bandit (MOMAB). Contrary to most previously proposed approaches in MOMAB, we do not search for the Pareto front, instead we aim for a fair solution, which is important for instance when each objective corresponds to the payoff of a different agent. To encode fairness, we use the Generalized Gini Index (GGI), a well-known criterion developed in economics. To optimize this criterion, we proposed a gradient-based algorithm that exploits the convexity of GGI. We evaluated our algorithm on two domains and obtained promising experimental results.

Several multi-objective reinforcement learning algorithm have been proposed in the literature (Gábor et al., 1998; Roijers et al., 2013). Most of these methods make use of a simple linear aggregation function. As a future work, it would be interesting to extend our work to the reinforcement learning setting, which would be useful to solve the electric battery control problem even more finely.

Acknowledgements

The authors would like to thank Vikram Bhattacharjee and Orkun Karabasoglu for providing the battery model. This research was supported in part by the European Communitys Seventh Framework Programme (FP7/2007-2013) under grant agreement 306638 (SUPREL).

References

Abernethy, Jacob D., Bartlett, Peter L., and Hazan, Elad. Blackwell approachability and no-regret learning are equivalent. In *COLT 2011 - The 24th Annual Conference on Learning Theory, June 9-11, 2011, Budapest, Hungary*, pp. 27–46, 2011.

Auer, P., Cesa-Bianchi, N., and Fischer, P. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47:235–256, 2002a.

Auer, P., Cesa-Bianchi, N., Freund, Y., and Schapire, R.E. The nonstochastic multiarmed bandit problem. *SIAM Journal on Computing*, 32(1):48–77, 2002b.

Azar, Yossi, Feige, Uriel, Feldman, Michal, and Tennenholtz, Moshe. Sequential decision making with vector outcomes. In *ITCS*, pp. 195–206, 2014.

Boyd, Stephen and Vandenberghe, Lieven. *Convex Optimization*. Cambridge University Press, New York, NY, USA, 2004.

Bubeck, S. and Cesa-Bianchi, N. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends in Machine Learning*, 5(1):1–122, 2012.

Buczolich, Zoltán and Székely, Gábor J. When is a weighted average of ordered sample elements a maximum likelihood estimator of the location parameter? *Advances in Applied Mathematics*, 10(4):439 – 456, 1989.

Dalton, H. The measurement of inequality of incomes. *Economic J.*, 30(348–361), 1920.

Dambrowski, J. Review on methods of state-of-charge estimation with viewpoint to the modern $LiFePO_4/Li_4Ti_5O_{12}$ lithium-ion systems. In *International Telecommunication Energy Conference*, 2013.

Drugan, M. M. and Nowé, A. Designing multi-objective multi-armed bandits algorithms: A study. In *IJCNN*, pp. 1–8, 2013.

Gábor, Zoltán, Kalmár, Zsolt, and Szepesvári, Csaba. Multi-criteria reinforcement learning. In *ICML*, pp. 197–205, 1998.

Gao, Lijun, Chen, Shenyi, and Dougal, Roger A. Dynamic lithium-ion battery model for system simulation. *IEEE Trans. on Components and Packaging Technologies*, 25 (3):495–505, 2002.

Hazan, Elad. *Introduction to Online Convex Optimization*. NOW, 2016.

Johnson, V.H. Battery performance models in ADVISOR. *Journal of Power Sources*, 110:312–329, 2002.

Lai, T. L. and Robbins, Herbert. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6(1):4–22, 1985.

Mahdavi, Mehrdad, Yang, Tianbao, and Jin, Rong. Stochastic convex optimization with multiple objectives. In *NIPS*, pp. 1115–1123. 2013.

Mannor, Shie, Tsitsiklis, John N., and Yu, Jia Yuan. Online learning with sample path constraints. *Journal of Machine Learning Research*, 10:569–590, 2009. doi: 10.1145/1577069.1577089.

Mannor, Shie, Perchet, Vianney, and Stoltz, Gilles. Approachability in unknown games: Online learning meets multi-objective optimization. In *Proceedings of The 27th Conference on Learning Theory, COLT 2014, Barcelona, Spain, June 13-15, 2014*, pp. 339–355, 2014.

Ogryczak, W. and Sliwinski, T. On solving linear programs with the ordered weighted averaging objective. *Eur. J. Operational Research*, 148:80–91, 2003.

Ogryczak, W., Perny, P., and Weng, P. On minimizing ordered weighted regrets in multiobjective Markov decision processes. In *ADT*, Lecture Notes in Artificial Intelligence, 2011.

Ogryczak, W., Perny, P., and Weng, P. A compromise programming approach to multiobjective Markov decision processes. *International Journal of Information Technology & Decision Making*, 12:1021–1053, 2013.

Pigou, A. *Wealth and Welfare*. Macmillan, 1912.

Press, W.H. Bandit solutions provide unified ethical models for randomized clinical trials and comparative effectiveness research. *PNAS*, 106(52):22398–22392, 2009.

Roijers, Diederik M., Vamplew, Peter, Whiteson, Shimon, and Dazeley, Richard. A survey of multi-objective sequential decision-making. *J. Artif. Intell. Res.*, 48: 67–113, 2013.

Shalev-Shwartz, Shai. Online learning and online convex optimization. *Foundations and Trends in Machine Learning*, 4(2):107–194, 2012.

Slivkins, A. Contextual bandits with similarity information. *J. Mach. Learn. Res.*, 15:2533–2568, 2014.

Steuer, R.E. and Choo, E.-U. An interactive weighted Tchebycheff procedure for multiple objective programming. *Mathematical Programming*, 26:326–344, 1983.

Tao, Gao. Research on $LiMn_2O_4$ battery's state of charge estimation with the consideration of degradation. Technical report, Tsinghua University, 2012.

Weymark, John A. Generalized gini inequality indices. *Mathematical Social Sciences*, 1(4):409 – 430, 1981.

Yager, R.R. On ordered weighted averaging aggregation operators in multi-criteria decision making. *IEEE Trans. on Syst., Man and Cyb.*, 18:183–190, 1988.

Zinkevich, Martin. Online convex programming and generalized infinitesimal gradient ascent. In *ICML*, 2003.