Online Sparse Coding with Bandit Feedback

Djallel Bouneffouf

Online Sparse Coding with Bandit Feedback

Djallel Bouneffouf
IBM Research
Djallel.Bouneffouf@IBM.com

Abstract

We consider a novel variant of the contextual bandit problem (i.e., the multi-armed bandit with side-information, or context, available to a decision-maker) where the reward associated with each context-based decision may not always be observed ("missing rewards"). This new problem is motivated by certain on-line settings including clinical trial and ad recommendation applications. In order to address the missing-reward setting, we propose to combine the standard contextual bandit approach with an unsupervised learning mechanism, such as, for example, sparse coding. Unlike standard contextual bandit methods, we are able to learn from all contexts, even those with missing rewards, by improving the representation of a context (via dictionary learning); when the reward is available, the standard contextual bandit learning mechanism is used. Promising empirical results are obtained on several real-life datasets.

1 Introduction

Sequential decison making is a common problem in many practical applications where the agent must choose the best action to perform at each iteration in order to maximize the cumulative reward over some period of time. One of the key challenges is achieve a good trade-off between the exploration of new actions and the exploitation of known actions. This exploration vs exploitation trade-off in sequential decision making problems is often formulated as the *multi-armed bandit (MAB)* problem: given a set of bandit "arms" (actions), each associated with a fixed but unknown reward probability distribution [63, 6, 44, 43, 3, 25, 40, 42, 16, 39, 17, 22, 48, 26, 45, 21, 31, 20, 23, 33, 18, 27, 28, 29, 4], an agent selects an arm to play at each iteration, and receives a reward, drawn according to the selected arm's distribution, independently from the previous actions.

A particularly useful version of MAB is the *contextual multi-armed bandit (CMAB)*, or simply the *contextual bandit* problem, where at each iteration, before choosing an arm, the agent observes an N-dimensional *context*, or *feature vector*. Over time, the goal is to learn the relationship between the context vectors and the rewards, in order to make better prediction which action to choose given the context [70, 8, 80, 9, 70, 79, 67, 55, 50, 17, 85, 72, 49, 89, 68, 1, 10, 76, 73, 82, 11, 71, 86] and anomaly detection [57].

For example, the contextual bandit approach is commonly used in various practical sequential decision problems with side information (context), from clinical trials [87] to recommender system [75], where the patient's information (medical history, etc.) or an online user's profile provide a context for making a better decision about the treatment to propose or an ad to show, and the reward represents the outcome of the selected action, such as, for example, success or failure of a particular treatment option.

In this paper, we consider a new problem setting, referred to as *contextual bandit with missing rewards*, where the agent may not always observe the reward, although the context is always observable. This setting is motivated by several real-life applications where the reward associated with a selected action can be missing (unobservable by an agent) for various reasons. For instance, in medical decision-making settings, the doctor can decide on a specific treatment option for a patient, but the patient may not come back for follow-up appointments; though the reward feedback regarding

31th Conference on Neural Information Processing Systems (NIPS 2017), Long Beach Convention Center.

the treatment success is missing, the context (patient's medical record) is still available and can be potentially used to learn more about the patient's population. A different example of missing rewards can occur in information retrieval or online search settings, where the user enters a search request, but, for various reasons, may not click on any of the suggested website links, and thus the reward feedback about those choices is missing. Yet another example can be online advertisement, where the user clicking on a proposed ad represents a positive reward, but the absence of a click can be interpreted either as negative reward (the user did not like the ad), or can be a consequence of a a bug or a connection loss.

The missing-reward contextual bandit framework proposed here aims to capture the situations described above, and provide an approach to always exploiting the context information in order to improve future decisions, even if some rewards are missing. More specifically, we will combine unsupervised learning via sparse coding (dictionary learning) with the standard contextual bandit: dictionary learning allows to learn good representations from all contexts, with and without the observed rewards, while the contextual bandit on top of sparse codes makes use of the reward information when it is available. We demonstrate on several real-life datasets that the proposed approach consistently outperforms the standard contextual bandit approach.

2 Related Work

The multi-armed bandit problem is a model of exploration versus exploitation trade-off, where a player gets to pick within a finite set of decisions the one maximizing the cumulative reward. This problem has been extensively studied. Optimal solutions have been provided using a stochastic formulation [46, 69, 66, 78, 35, 54, 14, 84, 64, 61, 34, 59, 41, 51, 36], a Bayesian formulation [83, 62, 30, 32, 47, 19, 38, 24, 37, 53, 52, 15], or using an adversarial formulation [5, 7]. However, these approaches do not take into account the context which may affect to the arm's performance. In LINUCB [65, 56] and in Contextual Thompson Sampling (CTS) [2], the authors assume a linear dependency between the expected reward of an action and its context; the representation space is modeled using a set of linear predictors. However, these algorithms assume that the agent can observe the reward at each iteration, which is not the case in many practical applications, including those discussed earlier in this paper.

Authors in [12] studies considering some kind of incomplete feedback called "Partial Monitoring (PM)", which is a general framework for sequential decision making problems with incomplete feedback that allows the learner, when it is possible, to retrieve the expected value of actions through an analysis of the feedback matrix, both of which are assumed to be known to the learner. In [58] authors study a variant of the stochastic multi-armed bandit (MAB) problem in which the rewards are corrupted. In this framework, motivated by privacy preserving in online recommender systems, the goal is to maximize the sum of the (unobserved) rewards, based on the observation of transformation of these rewards through a stochastic corruption process with known parameters. We can say that our setting is similare to the online semi-supervied learning [90, 77], which is a field of machine learning that studies learning from both labeled and unlabeled examples in an online setting. However in the their setting they receive the true label at each iteration, and we receive a bandit feedback.

3 Problem Setting

Algorithm 1 presents at a high-level the contextual bandit setting, where $c(t) \in C$ (we will assume here $C = \mathbf{R}^N$) is a vector describing the context at time t, $r_i(t) \in [0, 1]$ is the reward of the action i at time t, and $r(t) \in [0, 1]^K$ denotes a vector of rewards for all arms at time t. Also, $P_{c,r}$ denotes a joint probability distribution over (c, r), A denotes a set of K actions, $A = \{1, ..., K\}$, and $\pi : C \to A$ denotes a policy.

Algorithm 1 Contextual Bandit with missing rewards

1: **Repeat**
2: $(c(t), r(t))$ is drawn according to $D_{c,r}$
3: $c(t)$ is revealed to the player
4: The player chooses an action $i = \pi_t(c(t))$
5: The reward $r_i(t)$ is revealed with some probability
6: The player updates its policy π_t
7: $t = t + 1$
8: **Until** t=T

4 Our Approach: Bandit with Sparse-Coded Context

In order to make use of the context even in the absence of the corresponding reward, we propose to use an unsupervised learning approach; specifically, we use sparse coding as a representation learning step in contextual bandit problem and learn simultaneously, via alternating-minimization, the dictionary $\boldsymbol{D}$, codes $\{\boldsymbol{\alpha}_1, \cdots, \boldsymbol{\alpha}_n\}$ and the parameters $\boldsymbol{\theta}$ used to predict the expected reward for each arm. Our approach is described in Alg. 2. The main changes, as compared to the standard online dictionary learning algorithm of [74], are highlighted in Alg. 2.

At each iteration in Alg. 2, the next batch of samples is received and the corresponding codes, in the dictionary, are computed; next, we add k_n new dictionary elements sampled at random from $\mathbb{R}^m$ (i.e., k_n random linear projections of the input sample). The choice of the parameter k_n is important; one approach is to tune it (e.g., by cross-validation), while another is to adjust it dynamically, based on the dictionary performance: e.g., if the environment is changing, the old dictionary may not be able to represent the new input well, leading to decline in the representation accuracy, which triggers neurogenesis. Herein, we use as the performance measure the Pearson correlation between a new sample and its representation in the current dictionary $r(\boldsymbol{x}_t, \boldsymbol{D}^{(t-1)}\boldsymbol{\alpha}_t)$, i.e. denoted as $p_c(\boldsymbol{x}_t, \boldsymbol{D}^{(t-1)}, \boldsymbol{\alpha}_t)$ (for a batch of data, the average over $p_c(.)$ is taken). If it drops below a certain pre-specified threshold γ (where $0 \leq \gamma \leq 1$), the neurogenesis is triggered. The number k_n of new dictionary elements is proportional to the error $1 - p_c(\cdot)$, so that worse performance will trigger more neurogenesis, and vice versa; the maximum number of new elements is bounded by c_k.

We refer to this approach as *conditional neurogenesis* as it involves the *conditional birth* of new elements. Next, k_n random elements are generated and added to the current dictionary, and the memory matrices $\boldsymbol{A}, \boldsymbol{B}$ are updated, respectively, to account for larger dictionary. Finally, the sparse code is recomputed for $\boldsymbol{x}_t$ (or, all the samples in the current batch) with respect to the extended dictionary.

The next step is the dictionary update, which uses, similarly to the standard online dictionary learning, the block-coordinate descent approach. However, the objective function includes additional regularization terms:

$$\boldsymbol{D}^{(t)} = \arg\min_{\boldsymbol{D}\in\mathcal{C}} \frac{1}{t}\sum_{i=1}^{t} \frac{1}{2}||\boldsymbol{x}_i - \boldsymbol{D}\boldsymbol{\alpha}_i||_2^2 + \lambda_g \sum_j ||\boldsymbol{d}_j||_2 + \sum_j \lambda_j ||\boldsymbol{d}_j||_1. \tag{1}$$

The first term is the standard reconstruction error, as before. The second term, l_1/l_2-regularization, promotes group sparsity over the dictionary entries, where each group corresponds to a column, i.e. a dictionary element. The group-sparsity [88] regularizer causes some columns in $\boldsymbol{D}$ to be set to zero (i.e. the columns less useful for accurate data representation), thus effectively eliminating the corresponding dictionary elements from the dictionary ("killing" the corresponding hidden units). As it was mentioned previously, [13] used the l_1/l_2-regularizer in dictionary learning, though not in online setting, and without neurogenesis.

Finally, the third term imposes l_1-regularization on dictionary elements thus promoting sparse dictionary, besides the sparse coding. Introducing sparsity in dictionary elements, corresponding to the sparse connectivity of hidden units in the neural net representation of a dictionary, is motivated by both their biological plausibility (neuronal connectivity tends to be rather sparse in multiple brain networks), and by the computational advantages this extra regularization can provide, as we observe later in experiments.

Algorithm 2 Contextual Bandit with Dictionary Learning (CB-DL)

Require: Data stream $\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n \in \mathbb{R}^m$; initial dictionary $\boldsymbol{D} \in \mathbb{R}^{m \times k}$; l is the number of arms; number of non-zeros in a dictionary element, β_d; number of non-zeros in a code, β_c; Thompson-Sampling parameter v.

1: Initialize: $\boldsymbol{A} \leftarrow \boldsymbol{\epsilon}, \boldsymbol{g} \leftarrow \boldsymbol{\epsilon}$ % reset the "memory"
$\boldsymbol{B} = I_k, \hat{\mu} = 0_k, f = 0_k$ % Thompson Sampling parameters
2: **for** $t = 1$ to n **do**
3: Input $\boldsymbol{x}_t$
% Sparse coding:
4: $\boldsymbol{\alpha}_t = \arg_{\boldsymbol{\alpha} \in \mathbb{R}^k} \min \frac{1}{2}||\boldsymbol{x}_t - \boldsymbol{D}\boldsymbol{\alpha}||_2^2 + \lambda_c||\boldsymbol{\alpha}||_1$ % λ_c tuned to have β_c non-zeros in α_t
% Contextual bandit with α code as context
5: **for** $i = 1$ to l **do**
6: Sample $\boldsymbol{w}_i$ from the $N(\hat{\mu}_i, v^2 B_i^{-1})$ distribution.
7: **end for**
8: Play arm $j = \underset{i \in \{1,\ldots,l\}}{argmax}\, \boldsymbol{\alpha}_t^T \boldsymbol{w}_i$ and obtain r_t^i
9: $B_i = B_i + \boldsymbol{\alpha}_t \boldsymbol{\alpha}_t^T, f = f + \boldsymbol{\alpha}_t r_t^i, \hat{\mu_i} = B_i^{-1} f$ % End of Contextual Bandit
% Dictionary update through code
10: $\boldsymbol{\alpha}_t = \arg_{\boldsymbol{\alpha} \in \mathbb{R}^k} \min \frac{1}{2}||\boldsymbol{x}_t - \boldsymbol{D}\boldsymbol{\alpha}||_2^2 + \lambda_c||\boldsymbol{\alpha}||_1$ % Supervised re-optimization of code in the dictionary (with $\boldsymbol{\alpha}_t$ initialization)
% End of supervision
% "Memory" update:
11: $\boldsymbol{A} \leftarrow \boldsymbol{A} + \boldsymbol{\alpha}_t \boldsymbol{\alpha}_t^T$, $\boldsymbol{B} \leftarrow \boldsymbol{B} + \boldsymbol{x}_t \boldsymbol{\alpha}_t^T$
% Dictionary update by block-coordinate descent with l_1/l_2 group sparsity
12: **repeat**
13: **for** $j = 1$ to k **do**
14: $\boldsymbol{u}_j \leftarrow \frac{\boldsymbol{b}_j - \sum_{k \neq j} \boldsymbol{d}_k a_{jk}}{a_{jj}}$
15: % Sparsifying elements (optional):
$\boldsymbol{v}_j \leftarrow Prox_{\lambda_j||.||_1}(\boldsymbol{u}_j) = sgn(\boldsymbol{u}_j)(|\boldsymbol{u}_j| - \lambda_j)_+$, % λ_j tuned to get β_d non-zeros in $\boldsymbol{v}_j$
% Killing useless elements with l_1/l_2 group sparsity
16: $\boldsymbol{w}_j \leftarrow \boldsymbol{v}_j \left(1 - \frac{\lambda_g}{||\boldsymbol{v}_j||_2}\right)_+$
17: $\boldsymbol{d}_j \leftarrow \frac{\boldsymbol{w}_j}{\max(1, ||\boldsymbol{w}_j||_2)}$
18: **end for**
19: **until convergence**
20: **end for**
21: **return** $\boldsymbol{D}$

As in the original algorithm of [74], the above objective is optimized by the block-coordinate descent, where each block of variables corresponds to a dictionary element, i.e., a column in $\boldsymbol{D}$; the loop in steps 12-19 of the Alg. 2 iterates until convergence, defined by the magnitude of change between the two successive versions of the dictionary falling below some threshold. For each column update, the first and the last steps (the steps 14 and 17) are the same as in the original method of [74], while the two intermediate steps (the steps 15 and 16) are implementing additional regularization. Both steps 15 and 16 (sparsity and group sparsity regularization) are implemented using the standard proximal operators as described in [60]. Note that we actually use as input the desired number of non-zeros, and determine the corresponding sparsity parameter λ_c and λ_j using a binary search procedure. Overall, *the key features of our algorithm is the interplay of both the (conditional) birth and (group-sparsity) death of dictionary elements in an online setting.*

Theorem 1. *With probability $1 - \gamma$, where $0 \leq \gamma \leq 1$, we have the upper bound on the expected regret R(T) for the CB-DL (Algorithm 1) in the contextual bandit problem with K arms and d features (context size) is given as follows:*

$$E[R(T)] \leq \sqrt{8t}\sqrt{log[det([A+H]_t)] + dlog\lambda}(\sigma\sqrt{log[det([A+H]_t)] + dlog\lambda + 2log(\frac{1}{\delta})} + \frac{||\theta^*||}{\sqrt{\lambda}})$$

where $\Delta = \mu_k - \mu_k^e$ and σ the distance threshold, D_{max} a lipschitz constant and $0 \leq z \leq 1$ a constant parameter of the TS algorithm.

Proof. We consider the high probability event $\theta^* \in C_t$ for all $t > 0$.

$r_t = \pi_{t,a} - x_{t,a_t}^\top \mu^*$

$r_t = \pi_{t,a} - z_{t,a_t}^\top \mu^* + z_{t,a_t}^\top \mu^* - x_{t,a_t}^\top \mu^*$

$r_t \leq ||\pi_{t,a} - z_{t,a_t}^\top \mu^* + z_{t,a_t}^\top \mu^* - x_{t,a_t}^\top \mu^*||_2$

$r_t \leq ||\pi_{t,a} - z_{t,a_t}^\top \mu^*||_2 + ||z_{t,a_t}^\top \mu^* - x_{t,a_t}^\top \mu^*||_2$

where we adopt the cauchy-Schwarz inequality.

Now we investigate $||\pi_{t,a} - z_{t,a_t}^\top \mu^*||_2$ and $||z_{t,a_t}^\top \mu^* - x_{t,a_t}^\top \mu^*||_2$ separately.

Since μ_z^t is optimistic base.

$\pi_{t,a} - z_{t,a_t}^\top \mu_z^* \leq \langle z_{t,a_t}^\top \tilde{\mu}_z^t \rangle - \langle z_{t,a_t}^\top \tilde{\mu}_z^* \rangle \leq \langle z_{t,a_t}^\top \tilde{\mu}_z^t - \tilde{\mu}_z^* \rangle \leq \langle z_{t,a_t}^\top \tilde{\mu}_z^t - \hat{\mu}_z^t \rangle - \langle z_{t,a_t}^\top - z_{t,a_t}^\top \tilde{\mu}_z^* \rangle$

$r_t = \langle x_t^*, \theta^* \rangle - \langle x_t - \theta^* \rangle \leq \langle x_t, \tilde{\theta} \rangle - \langle x_t - \theta^* \rangle \; \theta^* \in C^*$

$= \langle x_t, \tilde{\theta} - \theta^* \rangle = \langle x_t, \hat{\theta} - \theta^* \rangle \; \langle x_t, \hat{\theta} - \hat{\theta} \rangle$ using Cauchy-Schwarz

$\leq ||x_t||_{(A+H)^{-1}} ||\hat{\theta} - \theta^*||_{(A+H)} + ||x_t||_{(A+H)^{-1}} ||\tilde{\theta} - \hat{\theta}||_{(A+H)}$

$\leq 2c_t ||x_t||_{(A+H)^{-1}} \; \theta^*, \tilde{\theta} \in C_t = \{\theta : ||\theta - \hat{\theta}_t||_{(A+H)} \leq c_t\}$

Since $x^\top \theta^* \in [-1, 1]$ for all $x \in X_t$ then we have $r_t \leq 2$. Therefore,

$r_t \leq min 2c_t ||x_t||_{(A+H)^{-1}}, 2 \leq 2c_t min c_t ||x_t||_{(A+H)^{-1}}, 1$

$r_t^2 \leq 4c_t^2 min c_t ||x_t||^2_{(A+H)^{-1}}, 1$

with $R_T = \sum_{t=1}^T r_t$ and using equation (1) we have,

$\sqrt{T \sum_{t=1}^T r_t^2} \leq \sqrt{T \sum_{t=1}^T 4c_t^2 min\{||x_t||^2_{(A+H)^{-1}}, 1\}}$

$\leq 2\sqrt{T} c_t \sqrt{\sum_{t=1}^T min\{||x_t||^2_{(A+H)^{-1}}, 1\}}$

since $x \leq 2log(1 + x)$ for $x \in [0, 1]$, we have

$\sum_{t=1}^T min\{||x_t||^2_{(A+H)^{-1}}, 1\} \leq 2 \sum_{t=1}^T log(1 + ||x_t||^2_{(A+H)^{-1}}) = 2(logdet(A + H)_t + dlog\lambda)$

□

5 Experiments

In this section we compare baseline method Thompson sampling which ignores the data with missing rewards to CB-DL- algorithm proposed in this paper. We consider Warfarin Problem [81]. Warfarin is an anticoagulant agent (Wysowski et al. 2007). Correctly dosing warfarin remains a significant challenge as the appropriate dosage is highly variable among individuals due to patient clinical, demographic and genetic factors. Physicians currently follow a fixed-dose strategy: they start patients on 5mg/day (the appropriate dose for the majority of patients) and slowly adjust the dose over the course of a few weeks by tracking the patient's anticoagulation levels. However, an incorrect initial dosage can result in highly adverse consequences such as stroke (if the initial dose is too low) or internal bleeding (if the initial dose is too high). Every year, nearly 43,000 emergency department visits in the United States are due to adverse events associated with inappropriate warfarin dosing (Budnitz et al. 2006). Thus, we tackle the problem of learning and assigning an appropriate initial dosage to patients by leveraging patient-specific factors.

So, in-order to evaluate our proposed algorithm, we take the dataset and remove the reword for some percentage of it. Which means that we allow the paper to see the bandit reward only in 1, 5 and 20 percent of the dataset. What we observe here is that, CTS with Dictionary update, is performing better than the baseline algorithm.

6 Conclusion

We consider a variant of the contextual bandit problem where the reward associated with each context-based decision may not always be observed. This problem is motivated by certain on-line settings including clinical trial and ad recommendation applications. We propose to combine the standard

Data	Supervision	Eval. Function	CTS	CTS with dictionary updates
Warfarin Dose				
5000DS	5%	Regret	0.6637 ± 0.0340	(0.5890 ± 0.0339)
3000DS	20%	Regret	0.6652 ± 0.0127	(0.5400 ± 0.0113)
3000DS	5%	Regret	0.6637 ± 0.0340	(0.5640 ± 0.0286)
3000DS	1%	Regret	0.6750 ± 0.0565	(0.6300 ± 0.0587)
2000DS	20%	Regret	0.6652 ± 0.0127	(0.6089 ± 0.0316)
2000DS	5%	Regret	0.6637 ± 0.0340	(0.6367 ± 0.0393)
2000DS	1%	Regret	0.6750 ± 0.0565	(0.6736 ± 0.0696)
1500DS	20%	Regret	0.6652 ± 0.0127	(0.5354 ± 0.0379)
1500DS	5%	Regret	0.6637 ± 0.0340	(0.5466 ± 0.0559)
1500DS	1%	Regret	0.6750 ± 0.0565	(0.6393 ± 0.0614)
1000DS	20%	Regret	0.6652 ± 0.0127	(0.5924 ± 0.0197)
1000DS	5%	Regret	0.6637 ± 0.0340	0.5905 ± 0.0376 (0.6062 ± 0.0467)
1000DS	1%	Regret	0.6750 ± 0.0565	(0.6136 ± 0.0610)
500DS	20%	Regret	0.6652 ± 0.0127	(0.5897 ± 0.0131)
500DS	5%	Regret	0.6637 ± 0.0340	(0.6013 ± 0.0309)
500DS	1%	Regret	0.6750 ± 0.0565	(0.6450 ± 0.0632)
300DS	20%	Regret	0.6652 ± 0.0127	(0.5988 ± 0.0162)
300DS	5%	Regret	0.6637 ± 0.0340	0.6443 ± 0.0286 (0.6208 ± 0.0260)
300DS	1%	Regret	0.6750 ± 0.0565	(0.6614 ± 0.0607)
200DS	5%	Regret	0.6637 ± 0.0340	(0.6585 ± 0.0214)
100DS	20%	Regret	0.6652 ± 0.0127	(0.6571 ± 0.0137)
100DS	5%	Regret	0.6637 ± 0.0340	(0.6552 ± 0.0240)
100DS	1%	Regret	0.6637 ± 0.0340	(0.6285 ± 0.0232)
30DS	5%	Regret	0.6637 ± 0.0340	(0.6342 ± 0.0364)
10DS	5%	Regret	0.6637 ± 0.0340	(0.4670 ± 0.0234)

Table 1: 1% data used for pre-training of dictionary in advance. 5 trials for all in poker, and 25 trials for all 5%. Batch size for dictionary is 200.

contextual bandit approach with sparse coding. Unlike standard contextual bandit methods, we are able to learn from all contexts, even those with missing rewards, by improving the representation of a context (via dictionary learning); when the reward is available, the standard contextual bandit learning mechanism is used. Promising empirical results are obtained on real-life datasets.

References

[1] Charu Aggarwal, Djallel Bouneffouf, Horst Samulowitz, Beat Buesser, Thanh Hoang, Udayan Khurana, Sijia Liu, Tejaswini Pedapati, Parikshit Ram, Ambrish Rawat, et al. How can ai automate end-to-end data science? *arXiv preprint arXiv:1910.14436*, 2019.

[2] Shipra Agrawal and Navin Goyal. Thompson sampling for contextual bandits with linear payoffs. In *ICML (3)*, pages 127–135, 2013.

[3] Robin Allesiardo, Raphaël Féraud, and Djallel Bouneffouf. A neural networks committee for the contextual bandit problem. In *International Conference on Neural Information Processing*, pages 374–381. Springer, Cham, 2014.

[4] Robin Allesiardo, Raphael Féraud, and Djallel Bouneffouf. Prise de décision contextuelle en bande organisée: Quand les bandits font un brainstorming. 2014.

[5] Peter Auer and Nicolò Cesa-Bianchi. On-line learning with malicious noise and the closure algorithm. *Ann. Math. Artif. Intell.*, 23(1-2):83–99, 1998.

[6] Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2-3):235–256, 2002.

[7] Peter Auer, Nicolò Cesa-Bianchi, Yoav Freund, and Robert E. Schapire. The nonstochastic multiarmed bandit problem. *SIAM J. Comput.*, 32(1):48–77, 2002.

[8] Avinash Balakrishnan, Djallel Bouneffouf, Nicholas Mattei, and Francesca Rossi. Using contextual bandits with behavioral constraints for constrained online movie recommendation. In *IJCAI*, pages 5802–5804, 2018.

[9] Avinash Balakrishnan, Djallel Bouneffouf, Nicholas Mattei, and Francesca Rossi. Incorporating behavioral constraints in online ai systems. In *AAAI*, 2019.

[10] Avinash Balakrishnan, Djallel Bouneffouf, Nicholas Mattei, and Francesca Rossi. Using multi-armed bandits to learn ethical priorities for online ai systems. *IBM Journal of Research and Development*, 63(4/5):1–1, 2019.

[11] Avinash Balakrishnan, Djallel Bouneffouf, Nicholas Mattei, and Francesca Rossi. Constrained decision-making and explanation of a recommendation, January 16 2020. US Patent App. 16/050,176.

[12] Gábor Bartók, Dean P Foster, Dávid Pál, Alexander Rakhlin, and Csaba Szepesvári. Partial monitoring—classification, regret bounds, and algorithms. *Mathematics of Operations Research*, 39(4):967–997, 2014.

[13] Samy Bengio, Fernando Pereira, Yoram Singer, and Dennis Strelow. Group sparse coding. In *Advances in Neural Information Processing Systems 22*. 2009.

[14] D Bouneffouf. Location-aware approach to improve context-based recommender system. *arXiv preprint arXiv:1303.0481*, 2013.

[15] Djallel Bouneffouf. Drars: un système de recommandation dynamique sensible au risque.

[16] Djallel Bouneffouf. Applying machine learning techniques to improve user acceptance on ubiquitous environement. *arXiv preprint arXiv:1301.4351*, 2013.

[17] Djallel Bouneffouf. *DRARS, a dynamic risk-aware recommender system*. PhD thesis, 2013.

[18] Djallel Bouneffouf. Evolution of the user's content: An overview of the state of the art. *arXiv preprint arXiv:1305.1787*, 2013.

[19] Djallel Bouneffouf. Exponentiated gradient linucb for contextual multi-armed bandits. *arXiv preprint arXiv:1305.2415*, 2013.

[20] Djallel Bouneffouf. Hybrid q-learning applied to ubiquitous recommender system. *arXiv preprint arXiv:1303.2651*, 2013.

[21] Djallel Bouneffouf. The impact of situation clustering in contextual-bandit algorithm for context-aware recommender systems. *arXiv preprint arXiv:1304.3845*, 2013.

[22] Djallel Bouneffouf. Improving adaptation of ubiquitous recommander systems by using reinforcement learning and collaborative filtering. *arXiv preprint arXiv:1303.2308*, 2013.

[23] Djallel Bouneffouf. Mobile recommender systems methods: An overview. *arXiv preprint arXiv:1305.1745*, 2013.

[24] Djallel Bouneffouf. Optimizing an utility function for exploration/exploitation trade-off in context-aware recommender system. *arXiv preprint arXiv:1303.0485*, 2013.

[25] Djallel Bouneffouf. Situation-aware approach to improve context-based recommender system. *arXiv preprint arXiv:1303.0481*, 2013.

[26] Djallel Bouneffouf. Towards user profile modelling in recommender system. *arXiv preprint arXiv:1305.1114*, 2013.

[27] Djallel Bouneffouf. Context-based information retrieval in risky environment. *arXiv preprint arXiv:1409.7729*, 2014.

[28] Djallel Bouneffouf. Etude des dimensions specifiques du contexte dans un systeme de filtrage d'informations. *arXiv preprint arXiv:1405.6287*, 2014.

[29] Djallel Bouneffouf. Freshness-aware thompson sampling. In *International Conference on Neural Information Processing*, pages 373–380. Springer, Cham, 2014.

[30] Djallel Bouneffouf. R-ucb: a contextual bandit algorithm for risk-aware recommender systems. *arXiv preprint arXiv:1408.2195*, 2014.

[31] Djallel Bouneffouf. Recommandation mobile, sensible au contexte de contenus\'evolutifs: Contextuel-e-greedy. *arXiv preprint arXiv:1402.1986*, 2014.

[32] Djallel Bouneffouf. Contextual bandit algorithm for risk-aware recommender systems. In *2016 IEEE Congress on Evolutionary Computation (CEC)*, pages 4667–4674. IEEE, 2016.

[33] Djallel Bouneffouf. Exponentiated gradient exploration for active learning. *Computers*, 5(1):1, 2016.

[34] Djallel Bouneffouf. Computing the dirichlet-multinomial log-likelihood function. *arXiv preprint arXiv:2007.11967*, 2020.

[35] Djallel Bouneffouf. Online learning with corrupted context: Corrupted contextual bandits. *arXiv preprint arXiv:2006.15194*, 2020.

[36] Djallel Bouneffouf, Charu Aggarwal, Horst Samulowitz, Beat Buesser, Thanh Hoang, Udayan Khurana, Sijia Liu, Tejaswini Pedapati, Parikshit Ram, Ambrish Rawat, et al. Survey on automated end-to-end data science.

[37] Djallel Bouneffouf and Inanc Birol. Ensemble minimum sum of squared similarities sampling for nyström-based spectral clustering. In *2016 International Joint Conference on Neural Networks (IJCNN)*, pages 3851–3855. IEEE, 2016.

[38] Djallel Bouneffouf and Inanc Birol. Theoretical analysis of the minimum sum of squared similarities sampling for nyström-based spectral clustering. In *2016 International Joint Conference on Neural Networks (IJCNN)*, pages 3856–3862. IEEE, 2016.

[39] Djallel Bouneffouf, Amel Bouzeghoub, and Alda Lopes Ganarski. Risk-aware recommender systems. In *International Conference on Neural Information Processing*, pages 57–65. Springer, Berlin, Heidelberg, 2013.

[40] Djallel Bouneffouf, Amel Bouzeghoub, and Alda Lopes Gançarski. Considering the high level critical situations in con-text-aware recommender systems. In *2nd International Workshop on Information Management for Mobile Applications*, page 26, 2012.

[41] Djallel Bouneffouf, Amel Bouzeghoub, and Alda Lopes Gançarski. contextual-bandit algorithm for context-aware recommender system. In *International conference on neural information processing*, pages 324–331. Springer, Berlin, Heidelberg, 2012.

[42] Djallel Bouneffouf, Amel Bouzeghoub, and Alda Lopes Gançarski. Exploration/exploitation trade-off in mobile context-aware recommender systems. In *Australasian Joint Conference on Artificial Intelligence*, pages 591–601. Springer, Berlin, Heidelberg, 2012.

[43] Djallel Bouneffouf, Amel Bouzeghoub, and Alda Lopes Gançarski. Following the user's interests in mobile context-aware recommender systems: The hybrid-e-greedy algorithm. In *2012 26th International Conference on Advanced Information Networking and Applications Workshops*, pages 657–662. IEEE, 2012.

[44] Djallel Bouneffouf, Amel Bouzeghoub, and Alda Lopes Gançarski. Hybrid-ε-greedy for mobile context-aware recommender system. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 468–479. Springer, Berlin, Heidelberg, 2012.

[45] Djallel Bouneffouf, Amel Bouzeghoub, and Alda Lopes Gançarski. Contextual bandits for context-based information retrieval. In *International Conference on Neural Information Processing*, pages 35–42. Springer, Berlin, Heidelberg, 2013.

[46] Djallel Bouneffouf and Emmanuelle Claeys. Hyper-parameter tuning for the contextual bandit. *arXiv preprint arXiv:2005.02209*, 2020.

[47] Djallel Bouneffouf and Raphael Féraud. Multi-armed bandit problem with known trend. *Neurocomputing*, 205:16–21, 2016.

[48] Djallel Bouneffouf, Romain Laroche, Tanguy Urvoy, Raphael Féraud, and Robin Allesiardo. Contextual bandit for active learning: Active thompson sampling. In *International Conference on Neural Information Processing*, pages 405–412. Springer, Cham, 2014.

[49] Djallel Bouneffouf, Srinivasan Parthasarathy, Horst Samulowitz, and Martin Wistub. Optimal exploitation of clustering and history information in multi-armed bandit. *arXiv preprint arXiv:1906.03979*, 2019.

[50] Djallel Bouneffouf and Irina Rish. A survey on practical applications of multi-armed and contextual bandits. In *The IEEE World Congress on Computational Intelligence (IEEE WCCI)*, 2020.

[51] Djallel Bouneffouf, Irina Rish, and Guillermo A Cecchi. Bandit models of human behavior.

[52] Djallel Bouneffouf, Irina Rish, and Guillermo A Cecchi. Bandit models of human behavior: Reward processing in mental disorders. In *International Conference on Artificial General Intelligence*, pages 237–248. Springer, Cham, 2017.

[53] Djallel Bouneffouf, Irina Rish, Guillermo A Cecchi, and Raphaël Féraud. Context attentive bandits: contextual bandit with restricted context. In *IJCAI 2017*, 2017.

[54] Djallel Bouneffouf, Sohini Upadhyay, and Yasaman Khazaeni. Contextual bandit with missing rewards. *arXiv preprint arXiv:2007.06368*, 2020.

[55] Anna Choromanska, Benjamin Cowen, Sadhana Kumaravel, Ronny Luss, Mattia Rigotti, Irina Rish, Paolo Diachille, Viatcheslav Gurev, Brian Kingsbury, Ravi Tejwani, et al. Beyond backprop: Online alternating minimization with auxiliary variables. In *International Conference on Machine Learning*, pages 1193–1202, 2019.

[56] Wei Chu, Lihong Li, Lev Reyzin, and Robert E. Schapire. Contextual bandits with linear payoff functions. In Geoffrey J. Gordon, David B. Dunson, and Miroslav Dudik, editors, *AISTATS*, volume 15 of *JMLR Proceedings*, pages 208–214. JMLR.org, 2011.

[57] Kaize Ding, Jundong Li, and Huan Liu. Interactive anomaly detection on attributed networks. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, WSDM '19, pages 357–365, New York, NY, USA, 2019. ACM.

[58] Pratik Gajane, Tanguy Urvoy, and Emilie Kaufmann. Corrupt bandits. *EWRL*, 2016.

[59] A Gupta, YS Ong, B Da, L Feng, and SD Handoko. 2016 ieee congress on evolutionary computation (cec). 2016.

[60] Rodolphe Jenatton, Julien Mairal, Guillaume Obozinski, and Francis Bach. Proximal methods for hierarchical sparse coding. *Journal of Machine Learning Research*, 2011.

[61] Andrew Teoh Beng Jin. *Neural Information Processing: 21st International Conference, ICONIP 2014, Kuching, Malaysia, November 3-6, 2014: Proceedings*. Springer, 2014.

[62] Emilie Kaufmann, Nathaniel Korda, and Rémi Munos. Thompson Sampling: An Asymptotically Optimal Finite Time Analysis. In *Algorithmic Learning Theory, Proc. of the 23rd International Conference (ALT)*, volume LNCS 7568, pages 199–213, Lyon, France, 2012. Springer.

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

[64] Chi Sing Leung. *Neural Information Processing: 19th International Conference, ICONIP 2012, Doha, Qatar, November 12-15, 2012, Proceedings, Part III*. Springer, 2012.

[65] Lihong Li, Wei Chu, John Langford, and Robert E. Schapire. A contextual-bandit approach to personalized news article recommendation. *CoRR*, 2010.

[66] Baihan Lin, Djallel Bouneffouf, and Guillermo Cecchi. Online learning in iterated prisoner's dilemma to mimic human behavior. *arXiv preprint arXiv:2006.06580*, 2020.

[67] Baihan Lin, Djallel Bouneffouf, Guillermo A Cecchi, and Irina Rish. Contextual bandit with adaptive feature extraction. In *2018 IEEE International Conference on Data Mining Workshops (ICDMW)*, pages 937–944. IEEE, 2018.

[68] Baihan Lin, Guillermo Cecchi, Djallel Bouneffouf, Jenna Reinen, and Irina Rish. Reinforcement learning models of human behavior: Reward processing in mental disorders. *arXiv preprint arXiv:1906.11286*, 2019.

[69] Baihan Lin, Guillermo Cecchi, Djallel Bouneffouf, Jenna Reinen, and Irina Rish. Unified models of human behavioral agents in bandits, contextual bandits and rl. *arXiv preprint arXiv:2005.04544*, 2020.

[70] Baihan Lin, Guillermo Cecchi, Djallel Bouneffouf, and Irina Rish. Adaptive representation selection in contextual bandit with unlabeled history. 2018.

[71] Baihan Lin, Guillermo A Cecchi, Djallel Bouneffouf, Jenna Reinen, and Irina Rish. A story of two streams: Reinforcement learning models from human behavior and neuropsychiatry. In *AAMAS*, pages 744–752, 2020.

[72] Sijia Liu, Parikshit Ram, Djallel Bouneffouf, Gregory Bramble, Andrew R Conn, Horst Samulowitz, and Alexander G Gray. Automated machine learning via admm. *CoRR, abs/1905.00424*, 2019.

[73] Sijia Liu, Parikshit Ram, Deepak Vijaykeerthy, Djallel Bouneffouf, Gregory Bramble, Horst Samulowitz, Dakuo Wang, Andrew Conn, and Alexander G Gray. An admm based framework for automl pipeline configuration. In *AAAI*, pages 4892–4899, 2020.

[74] Julien Mairal, Francis Bach, Jean Ponce, and Guillermo Sapiro. Online dictionary learning for sparse coding. In *Proceedings of the 26th annual international conference on machine learning*, 2009.

[75] Jérémie Mary, Romaric Gaudel, and Philippe Preux. Bandits and recommender systems. In *Machine Learning, Optimization, and Big Data - First International Workshop, MOD 2015*, pages 325–336, 2015.

[76] S Mehta, F Rossi, KR Varshney, A Balakrishnan, D Bouneffouf, N Mattei, R Noothigattu, R Chandra, P Madan, M Campbell, et al. Ai ethics. 2019.

[77] II Ororbia, G Alexander, C Lee Giles, and David Reitter. Online semi-supervised learning with deep hybrid boltzmann machines and denoising autoencoders. *arXiv preprint arXiv:1511.06964*, 2015.

[78] Parikshit Ram, Sijia Liu, Deepak Vijaykeerthi, Dakuo Wang, Djallel Bouneffouf, Greg Bramble, Horst Samulowitz, and Alexander G Gray. Solving constrained cash problems with admm. *arXiv preprint arXiv:2006.09635*, 2020.

[79] Matthew Riemer, Michele Franceschini, Djallel Bouneffouf, and Tim Klinger. Generative knowledge distillation for general purpose function compression. In *NIPS 2017 Workshop on Teaching Machines, Robots, and Humans*, volume 5, page 30, 2017.

[80] Matthew Riemer, Tim Klinger, Djallel Bouneffouf, and Michele Franceschini. Scalable recollections for continual lifelong learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 1352–1359, 2019.

[81] Ashkan Sharabiani, Adam Bress, Elnaz Douzali, and Houshang Darabi. Revisiting warfarin dosing using machine learning techniques. *Computational and mathematical methods in medicine*, 2015, 2015.

[82] Shubham Sharma, Yunfeng Zhang, Jesús M Ríos Aliaga, Djallel Bouneffouf, Vinod Muthusamy, and Kush R Varshney. Data augmentation for discrimination prevention and bias disambiguation. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 358–364, 2020.

[83] W.R. Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25:285–294, 1933.

[84] Kristina Toutanova and Hua Wu. Proceedings of the 52nd annual meeting of the association for computational linguistics (volume 1: Long papers). In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2014.

[85] Sohini Upadhyay, Mayank Agarwal, Djallel Bounneffouf, and Yasaman Khazaeni. A bandit approach to posterior dialog orchestration under a budget. *NIPS 2018*.

[86] KR Varshney, M Campbell, M Singh, and F Rossi. Teaching ai agents ethical values using reinforcement learning and policy orchestration.

[87] Sofía S Villar, Jack Bowden, and James Wason. 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):199, 2015.

[88] Ming Yuan and Yi Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2006.

[89] Mikhail Yurochkin, Sohini Upadhyay, Djallel Bouneffouf, Mayank Agarwal, and Yasaman Khazaeni. Online semi-supervised learning with bandit feedback. 2019.

[90] B. Yver. Online semi-supervised learning: Application to dynamic learning from radar data. In *2009 International Radar Conference "Surveillance for a Safer World" (RADAR 2009)*, pages 1–6, Oct 2009.