LEARNING EFFECTIVE EXPLORATION STRATEGIES FOR CONTEXTUAL BANDITS

Anonymous authors
Paper under double-blind review

ABSTRACT

In contextual bandits, an algorithm must choose actions given observed contexts, learning from a reward signal that is observed only for the action chosen. This leads to an exploration/exploitation trade-off: the algorithm must balance taking actions it already believes are good with taking new actions to potentially discover better choices. We develop a meta-learning algorithm, MÉLÉÉ, that learns an exploration policy based on simulated, synthetic contextual bandit tasks. MÉLÉÉ uses imitation learning against these simulations to train an exploration policy that can be applied to true contextual bandit tasks at test time. We evaluate on both a natural contextual bandit problem derived from a learning to rank dataset as well as hundreds of simulated contextual bandit problems derived from classification tasks. MÉLÉÉ outperforms seven strong baselines on most of these datasets by leveraging a rich feature representation for learning an exploration strategy.

1 INTRODUCTION

In a contextual bandit problem, an agent attempts to optimize its behavior over a sequence of rounds based on limited feedback (Kaelbling 1994; Auer 2003; Langford & Zhang 2008). In each round, the agent chooses an action based on a context (features) for that round, and observes a reward for that action but no others (§2). Contextual bandit problems arise in many real-world settings like online recommendations and personalized medicine. As in reinforcement learning, the agent must learn to balance exploitation (taking actions that, based on past experience, it believes will lead to high instantaneous reward) and exploration (trying actions that it knows less about).

In this paper, we present a meta-learning approach to automatically learn a good exploration mechanism from data. To achieve this, we use synthetic supervised learning data sets on which we can simulate contextual bandit tasks in an offline setting. Based on these simulations, our algorithm, MÉLÉÉ (MÉta LEarner for Exploration) learns a good heuristic exploration strategy that should ideally generalize to future contextual bandit problems. MÉLÉÉ contrasts with more classical approaches to exploration (like ϵ-greedy or LinUCB; see §6), in which exploration strategies are constructed by expert algorithm designers. These approaches often achieve provably good exploration strategies in the worst case, but are potentially overly pessimistic and are sometimes computationally intractable.

At training time (§3.2), MÉLÉÉ simulates many contextual bandit problems from fully labeled synthetic data. Using this data, in each round, MÉLÉÉ is able to counterfactually simulate what would happen under all possible action choices. We can then use this information to compute regret estimates for each action, which can be optimized using the AggreVaTe imitation learning algorithm (Ross & Bagnell 2014). Our imitation learning strategy mirrors that of the meta-learning approach of Bachman et al. (2017) in the active learning setting. We present a simplified, stylized analysis of the behavior of MÉLÉÉ to ensure that our cost function encourages good behavior (§4).

Empirically, we use MÉLÉÉ to train an exploration policy on only synthetic datasets and evaluate this policy on both a contextual bandit task based on a natural learning to rank dataset as well as three hundred simulated contextual bandit tasks (§5.2). We compare the trained policy to a number of alternative exploration algorithms, and show the efficacy of our approach (§5.3).

1 Code release: the code is available online https://www.dropbox.com/sh/dc3v8po5cbu8zaw/AACul1f_4c4w12xO1eW0RVZ0a?dl=0
2 Preliminaries: Contextual Bandits and Policy Optimization

Contextual bandits is a model of interaction in which an agent chooses actions (based on contexts) and receives immediate rewards for that action alone. For example, in a simplified news personalization setting, at each time step \( t \), a user arrives and the system must choose a news article to display to them. Each possible news article corresponds to an action \( a_t \), and the user corresponds to a context \( x_t \). After the system chooses an article \( a_t \) to display, it can observe, for instance, the amount of time that the user spends reading that article, which it can use as a reward \( r_t(a_t) \).

Formally, we largely follow the setup and notation of Agarwal et al. (2014). Let \( \mathcal{X} \) be an input space of contexts (users) and \( [K] = \{1, \ldots, K\} \) be a finite action space (articles). We consider the statistical setting in which there exists a fixed but unknown distribution \( \mathcal{D} \) over pairs \( (x, r) \in \mathcal{X} \times [0, 1]^K \), where \( r \) is a vector of rewards (for convenience, we assume all rewards are bounded in \( [0, 1] \)). In this setting, the world operates iteratively over rounds \( t = 1, 2, \ldots \). Each round \( t \):

1. The world draws \( (x_t, r_t) \sim \mathcal{D} \) and reveals context \( x_t \).
2. The agent (randomly) chooses action \( a_t \in [K] \) based on \( x_t \), and observes reward \( r_t(a_t) \).

The goal of an algorithm is to maximize the cumulative sum of rewards over time. Typically the primary quantity considered is the average regret of a sequence of actions \( a_1, \ldots, a_T \) to the behavior of the best possible function in a prespecified class \( \mathcal{F} \):

\[
\lambda(a_1, \ldots, a_T) = \max_{f \in \mathcal{F}} \frac{1}{T} \sum_{t=1}^{T} \left[ r_t(f(x_t)) - r_t(a_t) \right]
\]

(1)

An agent is call no-regret if its average regret is zero in the limit of large \( T \).

To produce a good agent for interacting with the world, we assume access to a function class \( \mathcal{F} \) and to an oracle policy optimizer for that function class. For example, \( \mathcal{F} \) may be a set of single layer neural networks mapping user features \( x \in \mathcal{X} \) to predicted rewards for actions \( a \in [K] \) with low expected regret. We consider \( \mathcal{F} \) be a finite action space (articles). For each tuple \( (x_t, a_t, r_t(a_t), p_t(a_t)) \in \mathcal{X} \times [0, 1] \times [0, 1] \), where \( p_t(a_t) \) is the probability that the agent chose action \( a_t \), and the full history of interaction is \( h_t = ((x_i, a_i, r_i(a_i), p_i(a_i)))_{i=1}^{t} \). The oracle policy optimizer, POLOPT, takes as input a history of user interactions and outputs an \( f \in \mathcal{F} \) with low expected regret.

A standard example of a policy optimizer is to combine inverse propensity scaling (IPS) with a regression algorithm (Dudik et al., 2011). Here, given a history \( h \), each tuple \( (x, a, r, p) \) in that history is mapped to a multiple-output regression example. The input for this regression example is the same \( x \); the output is a vector of \( K \) costs, all of which are zero except the \( a \)-th component, which takes value \( r/p \). This mapping is done for all tuples in the history, and a supervised learning algorithm on the function class \( \mathcal{F} \) is used to produce a low-regret regressor \( f \). This is the function returned by the policy optimizer. IPS, and other estimators that have lower-variance than IPS (such as the doubly-robust estimator), have the property of being unbiased. In experiments, we use the direct method (Dudik et al., 2011) largely for its simplicity, however, MÉLÉÉ is agnostic to the type of the estimator used by the policy optimizer.

3 Approach: Learning and Effective Exploration Strategy

In order to have an effective approach to the contextual bandit problem, one must be able to both optimize a policy based on historic data and make decisions about how to explore. The exploration/exploitation dilemma is fundamentally about long-term payoffs: is it worth trying something potentially suboptimal now in order to learn how to behave better in the future? A particularly simple and effective form of exploration is \( \epsilon \)-greedy: given a function \( f \) output by POLOPT, act according to \( f(x) \) with probability \( (1 - \epsilon) \) and act uniformly at random with probability \( \epsilon \). Intuitively, one would hope to improve on such a strategy by taking more (any!) information into account; for instance, basing the probability of exploration on \( f \)'s uncertainty. In this section, we describe MÉLÉÉ, first by describing how it operates at test time when applied to a new contextual bandit problem (§3.3.1), and then by describing how to train it using synthetic simulated contextual bandit problems (§5.2).
3.1 Test Time Behavior of Mélée

Our goal in this paper is to learn how to explore from experience. The training procedure for Mélée will use offline supervised learning problems to learn an exploration policy \( \pi \), which takes two inputs: a function \( f \in \mathcal{F} \) and a context \( x \), and outputs an action. In our example, \( f \) will be the output of the policy optimizer on all historic data, and \( x \) will be the current user. This is used to produce an agent which interacts with the world, maintaining an initially empty history buffer \( h \), as:

1. The world draws \( (x_t, r_t) \sim \mathcal{D} \) and reveals context \( x_t \).
2. The agent computes \( f_t \leftarrow \text{POLOpt}(h) \) and a greedy action \( \hat{a}_t = \pi(f_t, x_t) \).
3. The agent plays \( a_t = \hat{a}_t \) with probability \( 1 - \mu \), and \( a_t \) uniformly at random otherwise.
4. The agent observes \( r_t(a_t) \) and appends \( (x_t, a_t, r_t(a_t), p_t) \) to the history \( h \), where \( p_t = \mu/K \) if \( a_t \neq \hat{a}_t \); and \( p_t = 1 - \mu + \mu/K \) if \( a_t = \hat{a}_t \).

Here, \( f_t \) is the function optimized on the historical data, and \( \pi \) uses it and \( x_t \) to choose an action. Intuitively, \( \pi \) might choose to use the prediction \( f_t(x_t) \) most of the time, unless \( f_t \) is quite uncertain on this example, in which case \( \pi \) might choose to return the second (or third) most likely action according to \( f_t \). The agent then performs a small amount of additional \( \mu \)-greedy-style exploration: most of the time it acts according to \( \pi \) but occasionally it explores some more. In practice \( \mu = 5 \), we find that setting \( \mu = 0 \) is optimal in aggregate, but non-zero \( \mu \) is necessary for our theory \cite{Ross2015}.

Importantly, we wish to train \( \pi \) using one set of tasks (for which we have fully supervised data on which to run simulations) and apply it to wholly different tasks (for which we only have bandit feedback). To achieve this, we allow \( \pi \) to depend representationally on \( f_t \) in arbitrary ways: for instance, it might use features that capture \( f_t \)’s uncertainty on the current example (see \cite{Ross2015} for details). We additionally allow \( \pi \) to depend in a task-independent manner on the history (for instance, which actions have not yet been tried): it can use features of the actions, rewards and probabilities in the history but not depend directly on the contexts \( x \). This is to ensure that \( \pi \) only learns to explore and not also to solve the underlying task-dependent classification problem. Because \( \pi \) needs to learn to be task independent, we found that if \( f_t \)’s predictions were uncalibrated, it was very difficult for \( \pi \) to generalize well to unseen tasks. Therefore, we additionally allow \( \pi \) to depend on a very small amount of fully labeled data from the task at hand, which we use to allow \( \pi \) to calibrate \( f_t \)’s predictions. In our experiments we use only 30 fully labeled examples, but alternative approaches to calibrating \( f_t \) that do not require this data would be preferable.

3.2 Training Mélée by Imitation Learning

The meta-learning challenge is: how do we learn a good exploration policy \( \pi \)? We assume we have access to fully labeled data on which we can train \( \pi \); this data must include context/reward pairs, but where the reward for all actions is known. This is a weak assumption: in practice, we use purely synthetic data as this training data; one could alternatively use any fully labeled classification dataset \cite{Beygelzimer2009}. Under this assumption about the data, it is natural to think of \( \pi \)’s behavior as a sequential decision making problem in a simulated setting, for which a natural class of learning algorithms to consider are imitation learning algorithms \cite{Ross2010, Ross2014, Ross2011, Chang2015, Ross2015}.

Informally, at training time, Mélée will treat one of these synthetic datasets as if it were a contextual bandit dataset. At each time step \( t \), it will compute \( f_t \) by running POLOpt on the historical data, and then ask: for each action, what would the long time reward look like if I were to take this action. Because the training data for Mélée is fully labeled, this can be evaluated for each possible action, and a policy \( \pi \) can be learned to maximize these rewards. More formally, in imitation learning, we assume training-time access to an expert, \( \pi^* \), whose behavior we wish to learn to imitate at test-time. From this, we can define an optimal reference policy \( \pi^* \), which effectively “cheats” at training time by looking at the true labels. The learning problem is then to estimate \( \pi \) to have as similar behavior to \( \pi^* \) as possible, but without access to those labels.

The imitation learning algorithm we use is AggreVaTe \cite{Ross2014} (closely related to DAgger \cite{Ross2011}), and is instantiate for the contextual bandits meta-learning problem in Alg. 1. AggreVaTe learns to choose actions to minimize the cost-to-go of the expert rather than the

\[ \text{minimize} \quad \mathbb{E}[C_T(x_t, a_t)] \quad \text{subject to} \quad a_t = \pi(f_t, x_t) \]

Under review as a conference paper at ICLR 2020
Algorithm 1 MÉLÉÉ (supervised training sets \{S_m\}, hypothesis class \(\mathcal{F}\), exploration rate \(\mu\), number of validation examples \(N_{val}\), feature extractor \(\Phi\))

1: for round \(n = 1, 2, \ldots, N\) do
2: initialize meta-dataset \(D = \emptyset\), choose \(S\) at random from \(\{S_m\}\), and set history \(h_0 = \emptyset\)
3: partition and permute \(S\) randomly into train \(Tr\) and validation \(Val\) where |\(Val\)| = \(N_{val}\)
4: for round \(t = 1, 2, \ldots, |Tr|\) do
5: let \((x_t, r_t) = Tr_t\)
6: for each action \(a = 1, \ldots, K\) do
7: optimize \(f_{t,a} = \text{POLOPT}(\mathcal{F}, h_{t-1} \oplus (x_t, a, r_t(a), 1 - \frac{K-1}{K}\mu))\) on augmented history
8: roll-out: estimate \(\hat{c}_a\), the cost-to-go of \(a\), using \(r_t(a)\) and a roll-out policy \(\pi^{\text{out}}\) on \(f_{t,a}\)
9: end for
10: compute \(f_t = \text{POLOPT}(\mathcal{F}, h_{t-1}\)\)
11: aggregate \(D \leftarrow D \oplus (\Phi(f_t, x_t, h_{t-1}, Val), \langle \hat{c}_1, \ldots, \hat{c}_K \rangle)\)
12: roll-in: \(a_t \sim \frac{\mu}{K} 1_K + (1 - \mu) \pi_{n-1}(f_t, x_t)\) with probability \(p_t\), where \(1\) is the ones-vector
13: append history \(h_t \leftarrow h_{t-1} \oplus (x_t, a_t, r_t(a_t), p_t)\)
14: end for
15: update \(\pi_n = \text{LEARN}(D)\)
16: end for
17: return \(\{\pi_n\}_{n=1}^N\)

zero-one classification loss of mimicking its actions. On the first iteration AggreVaTe collects data by observing the expert perform the task, and in each trajectory, at time \(t\), explores an action \(a\) in state \(s\), and observes the cost-to-go \(Q\) of the expert after performing this action.

Following the AggreVaTe template, MÉLÉÉ operates in an iterative fashion, starting with an arbitrary \(\pi\) and improving it through interaction with an expert. Over \(N\) rounds, MÉLÉÉ selects random training sets and simulates the test-time behavior on that training set. The core functionality is to generate a number of states \((f_t, x_t)\) on which to train \(\pi\), and to use the supervised data to estimate the value of every action from those states. MÉLÉÉ achieves this by sampling a random supervised training set and setting aside some validation data from it (line 3). It then simulates a contextual bandit problem on this training data; at each time step \(t\), it tries all actions and “pretends” like they were appended to the current history (line 7) on which it trains a new policy and evaluates it’s roll-out value (line 8). This yields, for each \(t\), a new training example for \(\pi\), which is added to \(\pi\)’s training set (line 11); the features for this example are features of the classifier based on true history (line 10) (and possibly statistics of the history itself), with a label that gives, for each action, the corresponding cost-to-go of that action (the \(c_a\)s computed in line 8). MÉLÉÉ then must commit to a roll-in action to actually take; it chooses this according to a roll-in policy (line 12). MÉLÉÉ has no explicit “exploitation policy”, exploitation happens when \(\pi\) chooses the same action as \(f_t\), while exploration happens when it chooses a different action. In learning to explore, MÉLÉÉ simultaneously learns when to exploit.

Roll-in actions. The distribution over states visited by MÉLÉÉ depends on the actions taken, and in general it is good to have that distribution match what is seen at test time. This distribution is determined by a roll-in policy (line 12), controlled in MÉLÉÉ by exploration parameter \(\mu \in [0, 1]\). As \(\mu \to 1\), the roll-in policy approaches a uniform random policy; as \(\mu \to 0\), the roll-in policy becomes deterministic. When the roll-in policy does not explore, it acts according to \(\pi(f_t, \cdot)\).

Roll-out values. The ideal value to assign to an action (from the perspective of the imitation learning procedure) is that total reward (or advantage) that would be achieved in the long run if we took this action and then behaved according to our final learned policy. Unfortunately, during training, we do not yet know the final learned policy. Thus, a surrogate roll-out policy \(\pi^{\text{out}}\) is used instead. A convenient, and often computationally efficient alternative, is to evaluate the value assuming all future actions were taken by the expert (Langford & Zadrozny 2005) or Daumé et al. (2009) (Ross & Bagnell 2014). In our setting, at any time step \(t\), the expert has access to the fully supervised reward vector \(r_t\) for the context \(x_t\). When estimating the roll-out value for an action \(a\), the expert will return the true reward value for this action \(r_t(a)\) and we use this as our estimate for the roll-out value.
4 Theoretical Guarantees

We analyze MÉLÉE, showing that the no-regret property of AGGREGATE can be leveraged in our meta-learning setting for learning contextual bandit exploration. In particular, we first relate the regret of the learner in line 15 to the overall regret of \( \pi \). This will show that, if the underlying classifier improves sufficiently quickly, MÉLÉE will achieve sublinear regret. We then show that for a specific choice of underlying classifier (BANDITRON), this is achieved. MÉLÉE is an instantiation of AGGREGATE [Ross & Bagnell, 2014]; as such, it inherits AGGREGATE’s regret guarantees.

**Theorem 1** (Thm 2.2 of Ross & Bagnell [2014], adapted) After \( N \) rounds, if \textsc{learn} (line 15) is a no-regret algorithm, then as \( N \to \infty \), with probability 1, it holds that \( J(\pi) \leq J(\pi^*) + 2T\sqrt{K\epsilon_{\text{class}}(T)} \), where \( J(\cdot) \) is the reward of the exploration policy, \( \pi \) is the average policy returned, and \( \epsilon_{\text{class}}(T) \) is the average regret for each \( \pi_n \) accurately predicting \( \hat{\epsilon} \), where

\[
\epsilon_{\text{class}}(T) = \min_{\pi \in \Pi} \frac{1}{N} \sum_{t=1}^{N} \left[ Q^*_{T-t+1}(s, \pi) - \min_{a} Q^*_{T-t+1}(s, a) \right]
\]

is the empirical minimum expected cost-sensitive classification regret achieved by policies in the class \( \Pi \) on all the data over the \( N \) iterations of training when compared to the Bayes optimal regressor, for \( U(T) \) the uniform distribution over \( \{1, \ldots, T\} \), \( d^t_n \) the distribution of states at time \( t \) induced by executing policy \( \pi \), and \( Q^* \) the cost-to-go of the imitation learning expert.

Thus, achieving low regret at the problem of learning \( \pi \) on the training data it observes (“\( D^\pi \) in MÉLÉE”), i.e. \( \epsilon_{\text{class}}(T) \) is small, translates into low regret in the contextual-bandit setting. At first glance this bound looks like it may scale linearly with \( T \). However, the bound in Theorem 1 is dependent on \( \epsilon_{\text{class}}(T) \). Note however, that \( s \) is a combination of the context vector \( x \) and the classification function \( f \). As \( T \to \infty \), one would hope that \( f \) improves significantly and \( \epsilon_{\text{class}}(T) \) decays quickly. Thus, sublinear regret may still be achievable when \( f \) learns sufficiently quickly as a function of \( T \). For instance, if \( f \) is optimizing a strongly convex loss function, online gradient descent achieves a regret guarantee of \( O\left(\frac{\log T}{\mu} \right) \) [Hazan et al., 2016, Theorem 3.3], potentially leading to a regret for MÉLÉE of \( O\left(\sqrt{\log T}/\mu \right) \).

The above statement is informal (it does not take into account the interaction between learning \( f \) and \( \pi \)). However, we can show a specific concrete example: we analyze MÉLÉE’s test-time behavior when the underlying learning algorithm is BANDITRON. BANDITRON is a variant of the multiclass Perceptron that operates under bandit feedback. Details of this analysis (and proofs, which directly follow the original BANDITRON analysis) are given in Appendix A; here we state the main result. Let \( \gamma_t = \Pr[r_t(f_t, x_t) = 1|x_t] - \Pr[r_t(f_\pi(x_t)) = 1|x_t] \), be the edge of \( f_t \) over \( f \), and let \( \Gamma = \frac{1}{T} \sum_{t=1}^{T} \gamma_t \) be an overall measure of the edge. For instance if \( \pi \) simply returns \( f \)’s prediction, then all \( \gamma_t = 0 \) and \( \Gamma = 1 \). We can then show the following:

**Theorem 2** Assume that for the sequence of examples, \( (x_1, r_1), (x_2, r_2), \ldots, (x_T, r_T) \), we have, for all \( t, ||x_t|| \leq 1 \). Let \( W^* \) be any matrix, let \( L \) be the cumulative hinge-loss of \( W^* \), let \( \mu \) be a uniform exploration probability, and let \( D = 2 ||W^*||_F^2 \) be the complexity of \( W^* \). Assume that \( \mathbb{E}\gamma_t \geq 0 \) for all \( t \). Then the number of mistakes \( M \) made by MÉLÉE with BANDITRON as POLOPT satisfies:

\[
E M \leq L + K\mu T + 3\max\left\{ D\Gamma/\mu, \sqrt{D\Gamma KT\mu}\right\} + \sqrt{D\Gamma T/\mu}
\]

where the expectation is taken with respect to the randomness of the algorithm.

Note that under the assumption \( \mathbb{E}\gamma_t \geq 0 \) for all \( t \), we have \( \Gamma \leq 1 \). The analysis gives the same mistake bound for BANDITRON but with the factor of \( \Gamma \), hence this result improves upon the BANDITRON analysis only when \( \Gamma < 1 \) (in the realizable setting, the number of mistakes is analogous to the regret). This result is highly stylized and the assumption that \( \mathbb{E}\gamma_t \geq 0 \) is overly strong. This assumption ensures that \( \pi \) never decreases the probability of a “correct” action. It does, however, help us understand the behavior of MÉLÉE, qualitatively: First, the quantity that matters in Theorem 2 \( \mathbb{E}\gamma_t \) is (in the 0/1 loss case) exactly what MÉLÉE is optimizing: the expected improvement for choosing an action against \( f_t \)’s recommendation. Second, the benefit of using \( \pi \) within BANDITRON is a local benefit: because \( \pi \) is trained with expert rollouts, as discussed in [14], the primary improvement in the analysis is to ensure that \( \pi \) does a better job predicting (in a single step) than
that the performance of M\^{}\text{\textregistered} textural bandit problems. We evaluate M\^{}\text{\textregistered} and perform thorough evaluation on reward of the algorithm. Specifically, to evaluate the performance of an exploration algorithm A for evaluation, we use progressive validation (Blum et al., 1999), which is exactly computing the distribution; the underlying classifier f is a linear model trained with a policy optimizer that runs stochastic gradient descent (details are in Appendix A.2). We seek to answer two questions experimentally: (1) How does M\^{}\text{\textregistered} compare empirically to alternative (expert designed) exploration strategies? (2) How important are the additional features used by M\^{}\text{\textregistered} in comparison to using calibrated probability predictions from f as features?

5 Experimental Setup and Results

Using a collection of synthetically generated classification problems, we train an exploration policy \(\pi\) using MÉLÉÉ \(\text{[Alg 1]}\). This exploration policy learns to explore on the basis of calibrated probabilistic predictions from f together with a predefined set of exploration features \(\text{(5.1)}\). Once \(\pi\) is learned and fixed, we follow the test-time behavior described in \(\text{(3.1)}\) to evaluate \(\pi\) on a set of contextual bandit problems. We evaluate MÉLÉÉ on a natural learning to rank task \(\text{(5.3.1)}\). To ensure that the performance of MÉLÉÉ generalizes beyond this single learning to rank task, we additionally perform thorough evaluation on 300 “simulated” contextual bandit problems, derived from standard classification tasks \(\text{(5.3.2)}\).

In all cases, the underlying classifier f is a linear model trained with a policy optimizer that runs stochastic gradient descent (details are in Appendix A.2). We seek to answer two questions experimentally: (1) How does MÉLÉÉ compare empirically to alternative (expert designed) exploration strategies? (2) How important are the additional features used by MÉLÉÉ in comparison to using calibrated probability predictions from f as features?

5.1 Training Details for the Exploration Policy

Exploration Features. In our experiments, the exploration policy is trained based on features \(\Phi\) (Alg 1 line 11). These features are allowed to depend on the current classifier \(f_t\), and on any part of the history except the inputs \(x_t\) in order to maintain task independence. We additionally ensure that its features are independent of the dimensionality of the inputs, so that \(\pi\) can generalize to datasets of arbitrary dimensions. The specific features we use are listed below; these are largely inspired by Konyushkova et al. (2017) but adapted and augmented to our setting.

The features of \(f_t\) that we use are: a) predicted probability \(p(a_t|f_t, x_t)\), we use a softmax over the predicted rewards from \(f_t\) to convert them to probabilities; b) entropy of the predicted probability distribution; c) a one-hot encoding for the predicted action \(f_t(x_t)\). The features of \(h_{t-1}\) that we use are: a) current time step \(t\); b) normalized counts for all previous actions predicted so far; c) average observed rewards for each action; d) empirical variance of the observed rewards for each action in the history. In our experiments, we found that it is essential to calibrate the predicted probabilities of the classifier \(f_t\). We use a very small held-out dataset, of size 30, to achieve this. We use Platt’s scaling (Platt, 1999; Lin et al., 2007) method to calibrate the predicted probabilities. Platt’s scaling works by fitting a logistic regression model to the classifier’s predicted scores.

Training Datasets. In our experiments, we follow Konyushkova et al. (2017) (and also Peters et al., 2014, in a different setting) and train the exploration policy \(\pi\) only on synthetic data. This is possible because the exploration policy \(\pi\) never makes use of \(x\) explicitly and instead only accesses it via \(f_t\)’s behavior on it. We generate datasets with uniformly distributed class conditional distributions. The datasets are always two-dimensional. Details are in Appendix A.1.

5.2 Evaluation Methodology

For evaluation, we use progressive validation (Blum et al. 1999), which is exactly computing the reward of the algorithm. Specifically, to evaluate the performance of an exploration algorithm \(\mathcal{A}\) on a dataset \(S\) of size \(n\), we compute the progressive validation return \(G(\mathcal{A}) = \frac{1}{n} \sum_{t=1}^{n} r_t(a_t)\) as the average reward up to \(n\), where \(a_t\) is the action chosen by the algorithm \(\mathcal{A}\) and \(r_t\) is the true reward. Because our evaluation is over 300 datasets, we report aggregate results in two forms. The simpler one is Win/Loss Statistics: We compare two exploration methods on a given dataset by counting the number of statistically significant wins and losses. An exploration algorithm \(\mathcal{A}\) wins over another algorithm \(B\) if the progressive validation return \(G(\mathcal{A})\) is statistically significantly larger than \(B\)’s return \(G(B)\) at the 0.01 level using a paired sample t-test. We also report cumulative distributions of rewards for each algorithm, following Zhang et al. (2019). In particular, for a given relative reward value \((x \in [0, 1])\), the corresponding CDF value for a given algorithm is the fraction of datasets on which this algorithm achieved reward at least \(x\). We compute relative reward by Min-
Max normalization: linearly transforming reward $y$ to $y' = \frac{y - \min_{y} \max_{y}}{\max_{y} - \min_{y}}$, where min & max are the minimum & maximum rewards among all exploration algorithms.

In our experiments, we compare to the following baseline exploration methods, keeping the policy optimization method fixed (details in Appendix A.4): $\epsilon$-greedy (Sutton, 1996); $\epsilon$-decreasing (Sutton & Barto, 1998); EG $\epsilon$-greedy (Li et al., 2010b); $\tau$-first. We additionally compare to three state-of-the-art exploration methods: LinUCB (Li et al., 2010a). Cover (Agarwal et al., 2014); and its variant Cover-NU (Bietti et al., 2018). We select the best hyperparameters following Bietti et al. (2018).

5.3 Experimental Results

5.3.1 Learning to Rank

We evaluate MÉLÉE on a natural learning to rank dataset. The dataset we consider is the Microsoft Learning to Rank dataset, variant MSLR-10K from Qin & Liu (2013). The dataset consists of feature vectors extracted from query-url pairs along with relevance judgment labels. The relevance judgments are obtained from a retired labeling set of a commercial web search engine (Microsoft Bing), which take 5 values from 0 (irrelevant) to 4 (perfectly relevant) and we drop the queries not labelled as any of the two extremes. In our experiments, we limit the number of labels to the two extremes: 0 and 4. A query-url pair is represented by a 136-dimensional feature vector. The dataset is highly imbalanced as the number of irrelevant queries is much larger than the number of relevant ones. To address this, we sample the number of irrelevant queries to match that of the relevant ones. To avoid correlations between the observed query-url pairs, we group the queries by the query ID, and sample a single query from each group. We convert relevance scores to losses with 0 indicating a perfectly relevant document, and 1 an irrelevant one.

Figure 1 shows the evaluation results on a subset of the MSLR-10K dataset. Since the performance is closely matched between the different exploration algorithms, we repeat the experiment 16 times with randomly shuffled permutations of the MSLR-10K dataset. Figure 1 (left) shows the learning curve of the trained policy $\pi$ as well as the baselines. Here, we see that MÉLÉE quickly achieves high reward, after about 100 examples the two strongest baselines catch up. By 200 examples all approaches have asymptoted. We exclude LinUCB from these runs because the required matrix inversions made it too computationally expensive. Figure 1 shows statistically-significant win/loss differences for each of the algorithms, across these 16 shuffles. Each row/column entry shows the number of times the row algorithm won against the column, minus the number of losses. MÉLÉE is the only algorithm that always wins more than it loses against other algorithms, and outperforms the nearest competition ($\epsilon$-decreasing) by 3 points.

5.3.2 Simulated Contextual Bandit Tasks

We additionally perform an exhaustive evaluation on simulated contextual bandit tasks to ensure that the performance of MÉLÉE generalizes beyond learning to rank. Following Bietti et al. (2018), we use a collection of 300 binary classification datasets from openml.org for evaluation; the precise list and download instructions is in Appendix A.3. These datasets cover a variety of different domains including text & image processing, medical, and sensory data. We convert classification datasets into cost-sensitive classification problems by using a 0/1 encoding. Given these fully supervised cost-sensitive multi-class datasets, we simulate the contextual bandit setting by only revealing the reward for the selected actions.

In Figure 2 (left), we show a representative learning curve. Here, we see that as more data becomes available, all the approaches improve (except $\tau$-first, which has ceased to learn after 2% of the data). MÉLÉE, in particular, is able to very quickly achieve near-optimal performance (in around 40 examples) in comparison to the best baseline which takes at least 200. In Figure 2 (right), we show the CDFs for the different algorithms. To help read this, at $x = 1.0$, MÉLÉE has a relative reward at least 1.0 on more than 40% of datasets, while $\epsilon$-decreasing and $\epsilon$-greedy achieve this on about 30% of datasets. We find that the two strongest baselines are $\epsilon$-decreasing and $\epsilon$-greedy (better when reward differences are small, toward the left of the graph). The two curves for $\epsilon$-decreasing and $\epsilon$-greedy coincide. This happens because the exploration probability $\epsilon_0$ for $\epsilon$-decreasing decays rapidly
To understand more directly how MÉLÉE compares to \( \epsilon \)-decreasing in low reward regimes (left of graph), we show a scatter plot of rewards achieved by MÉLÉE (x-axis) and \( \epsilon \)-decreasing (y-axis) on each of the 300 datasets, with statistically significant differences highlighted in red and insignificant differences in blue. Points below the diagonal line correspond to better performance by MÉLÉE (147 datasets).
Figure 3: Win/Loss counts for all pairs of algorithms (columns match the rows).

Figure 4: Scatterplots comparing MÉLÉÉ to the best baseline and to a variant with fewer features.

and points above to ϵ-decreasing (124 datasets). The remaining 29 had no statistically significant difference.

Finally, we consider the effect that the additional features have on MÉLÉÉ’s performance. In particular, we consider a version of MÉLÉÉ with all features (this is the version used in all other experiments) with an ablated version that only has access to the (calibrated) probabilities of each action from the underlying classifier f. The comparison is shown as a scatter plot in Figure 4 (right). Here, we can see that the full feature set does provide lift over just the calibrated probabilities, with a win-minus-loss improvement of 24 by adding additional features from which to learn to explore.

6 RELATED WORK AND DISCUSSION

The field of meta-learning is based on the idea of replacing hand-engineered learning heuristics with heuristics learned from data. One of the most relevant settings for meta-learning to ours is active learning, in which one aims to learn a decision function to decide which examples, from a pool of unlabeled examples, should be labeled. Past approaches to meta-learning for active learning include reinforcement learning-based strategies (Woodward & Finn, 2017; Fang et al., 2017), imitation learning-based strategies (Bachman et al., 2017), and batch supervised learning-based strategies (Konyushkova et al., 2017). Similar approaches have been used to learn heuristics for optimization (Li & Malik, 2016; Andrychowicz et al., 2016), multiarm (non-contextual) bandits (Maes et al., 2012), and neural architecture search (Zoph & Le, 2016), recently mostly based on (deep) reinforcement learning. While meta-learning for contextual bandits is most similar to meta-learning for
active learning, there is a fundamental difference that makes it significantly more challenging: in active learning, the goal is to select as few examples as you can to learn, so by definition the horizon is short; in contextual bandits, learning to explore is fundamentally a long-horizon problem, because what matters is not immediate reward but long term learning.

In reinforcement learning, Gupta et al. (2018) investigated the task of meta-learning an exploration strategy for a distribution of related tasks by learning a latent exploration space. Similarly, Xu et al. (2018) proposed a teacher-student approach for learning to do exploration in off-policy reinforcement learning. While these approaches are effective if the distribution of tasks is very similar and the state space is shared among different tasks, they fail to generalize when the tasks are different. Our approach targets an easier problem than exploration in full reinforcement learning environments, and can generalize well across a wide range of different tasks with completely unrelated features spaces.

There has also been a substantial amount of work on constructing “good” exploration policies, in problems of varying complexity: traditional bandit settings (Karnin & Anava, 2016), contextual bandits (Fraud et al., 2016) and reinforcement learning (Osband et al., 2016). In both bandit settings, most of this work has focused on the learning theory aspect of exploration: what exploration distributions guarantee that learning will succeed (with high probability)? MËLÉE, lacks such guarantees: in particular, if the data distribution of the observed contexts (\(\phi(f_t)\)) in some test problem differs substantially from that on which MËLÉE was trained, we can say nothing about the quality of the learned exploration. Nevertheless, despite fairly substantial distribution mismatch (synthetic \(\rightarrow\) real-world), MËLÉE works well in practice, and our stylized theory \(\S\) suggests that there may be an interesting avenue for developing strong theoretical results for contextual bandit learning with learned exploration policies, and perhaps other meta-learning problems.

In conclusion, we presented MËLÉE, a meta-learning algorithm for learning exploration policies in the contextual bandit setting. MËLÉE enjoys no-regret guarantees, and empirically it outperforms alternative exploration algorithm in most settings. One limitation of MËLÉE is the computational resources required during the offline training phase on the synthetic datasets. In the future, we will work on improving the computational efficiency for MËLÉE in the offline training phase and scale the experimental analysis to problems with larger number of classes.

REFERENCES


Sham M. Kakade, Shai Shalev-Shwartz, and Ambuj Tewari. Efficient bandit algorithms for online multiclass prediction. In ICML, 2008.


Wei Li, Xuerui Wang, Ruofei Zhang, Ying Cui, Jianchang Mao, and Rong Jin. Exploitation and exploration in a performance based contextual advertising system. In Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD ’10, pp. 27–36, New York, NY, USA, 2010b. ACM.


Supplementary Material For:
Learning Effective Exploration Strategies for Contextual Bandits

A STYLIZED TEST-TIME ANALYSIS FOR BANDITRON: DETAILS

The BANDITRONMÉLÉE algorithm is specified in Alg 2. The only difference between the sampled distribution and the original BANDITRON is the introduction of \( \pi \) in the sampling distribution. The original algorithm achieves the following mistake bound shown below, which depends on the notion of multi-class hinge-loss. The hinge-loss of \( \pi \) takes as arguments: \( \pi \) (the learned classification policy) and \( \mu \in (0, 1/(2K)) \) an added uniform exploration parameter. The BANDITRON assigns a multi-class classifier parameterized by a weight matrix of size \( K \times D \), where \( D \) is the input dimensionality. The BANDITRON assumes a pure multi-class setting in which the reward for one (“correct”) action is 1 and the reward for all other actions is zero.

At each round \( t \), a prediction \( \hat{a}_t \) is made according to \( f_t \) (summarized by \( W^t \)). We then define an exploration distribution that “most of the time” acts according to \( \pi(f_t, \cdot) \), but smooths each action with \( \mu \) probability. The chosen action \( a_t \) is sampled from this distribution and a binary reward is observed. The weights of the updated BANDITRON are updated according to the BANDITRON update rule using \( U^t \).

**Algorithm 2 BANDITRONMÉLÉE \((g, \mu)\)**

1: initialize \( W^1 = 0 \in \mathbb{R}^{K \times D} \)
2: for rounds \( t = 1 \ldots T \): do
3: observe \( x_t \in \mathbb{R}^D \)
4: compute \( \hat{a}_t = f_t(x_t) = \arg \max_{a \in K} (W^t x_t)_a \)
5: define \( Q^\mu(a) = \mu + (1 - K\mu)1[a = \pi(W^t, x_t)] \)
6: sample \( a_t \sim Q^\mu \)
7: observe reward \( r_t(a_t) \in \{0, 1\} \)
8: define \( \tilde{U}^t \in \mathbb{R}^{K \times D} \) as:
\[
\tilde{U}^t_{a_t} = x_t \left( \frac{1[\hat{a}_t = a] 1[a_t = a]}{Q^\mu(a)} - 1[\hat{a}_t = a] \right)
\]
9: update \( W^{t+1} = W^t + \tilde{U}^t \)
10: end for

The only difference between BANDITRONMÉLÉE and the original BANDITRON is the introduction of \( \pi \) in the sampling distribution. The original algorithm achieves the following mistake bound shown below, which depends on the notion of multi-class hinge-loss. In particular, the hinge-loss of \( W \) on \((x, r)\) is \( \ell(W, (x, r)) = \max_{a \neq a^*} \max\{0, 1 - (wx)_{a^*} + (wx)_a\} \), where \( a^* \) is the \( a \) for which \( r(a) = 1 \). The overall hinge-loss \( L \) is the sum of \( \ell \) over the sequence of examples.

**Theorem 3 (Thm. 1 and Corr. 2 of [Kakade et al. (2008)])** Assume that for the sequence of examples, \((x_1, r_1), (x_2, r_2), \ldots, (x_T, r_T)\), we have, for all \( t \), \( ||x_t|| \leq 1 \). Let \( W^* \) be any matrix, let \( L \) be the cumulative hinge-loss of \( W^* \), and let \( D = 2 ||W^*||_F^2 \) be the complexity of \( W^* \). The number of mistakes \( M \) made by the BANDITRON satisfies

\[
EM \leq L + K\mu T + 3 \max \left\{ \frac{D}{\mu}, \sqrt{DTK\mu} \right\} + \sqrt{DL/\mu}
\]

where the expectation is taken with respect to the randomness of the algorithm. Furthermore, in a low noise setting (there exists \( W^* \) with fixed complexity \( D \) and loss \( L \leq O(\sqrt{DT}) \), then by setting \( \mu = \sqrt{D/(T K)} \), we obtain \( EM \leq O(\sqrt{DT}) \).

We can prove an analogous result for BANDITRONMÉLÉE. The key quantity that will control how much \( \mu \) improves the execution of BANDITRONMÉLÉE is how much \( \pi \) improves on \( f_t \) when \( f_t \) is wrong. In particular, let \( \gamma_t = \text{Pr}[r_t(\pi(f_t, x_t) = 1) | x_t] - \text{Pr}[r_t(f_t(x_t)) = 1 | x_t] \) be the edge of \( \pi(f_t, \cdot) \) over \( f_t \), and let \( \Gamma = \sum_{t=1}^T \frac{1}{1 + \gamma_t} \) be an overall measure of the edge. (If \( \pi \) does nothing, then all \( \gamma_t = 0 \) and \( \Gamma = 1 \).) Given this quantity, we can prove the following

**Proof:** [Sketch] The proof is a small modification of the original proof of [Theorem 3]. The only change is that in the original proof, the following bound is used: \( E_t ||U^t||^2 / ||x_t||^2 = 1 + 1/\mu \leq 2/\mu \).
We use, instead: $E_t[\tilde{U}^t]^2 / ||x_t||^2 \leq 1 + E_t[\frac{1}{\mu + \gamma_t}] \leq \frac{2E_t[\frac{1}{\mu + \gamma_t}]}{\mu}$. The rest of the proof goes through identically.

A.1 Details of Synthetic Datasets

We generate datasets with uniformly distributed class conditional distributions. We generate 2D datasets by first sampling a random variable representing the Bayes classification error. The Bayes error is sampled uniformly from the interval $0.0$ to $0.5$. Next, we generate a balanced dataset where the data for each class lies within a unit rectangle and sampled uniformly. We overlap the sampling rectangular regions to generate a dataset with the desired Bayes error selected in the first step.

A.2 Implementation Details

Our implementation is based on scikit-learn (Pedregosa et al., 2011). We fix the training time exploration parameter $\mu$ to 0.1. We train the exploration policy $\pi$ on 82 synthetic datasets each of size 3000 with uniform class conditional distributions, a total of $246k$ samples ($\{A.1\}$). We train $\pi$ using a linear classifier [Breiman (2001)] and set the hyper-parameters for the learning rate, and data scaling methods using three-fold cross-validation on the whole meta-training dataset. For the classifier class $F$, we use a linear model trained with stochastic gradient descent. We standardize all features to zero mean and unit variance, or scale the features to lie between zero and one. To select between the two scaling methods, and tune the classifier’s learning rate, we use three-fold cross-validation on a small fully supervised training set of size 30 samples. The same set is used to calibrate the predicted probabilities of $f_k$.

A.3 List of Datasets

The list of datasets we can accessed at https://www.openml.org/d/<id> (799, 1000) (806, 1000) (845, 1000) (849, 1000) (866, 1000) (903, 1000) (907, 1000) (908, 1000) (909, 1000) (913, 1000) (917, 1000) (741, 1024) (1444, 1043) (1453, 1077)
Our experiments aim to determine how Mêlée compares to other standard exploration strategies. In particular, we compare to:

**ε-greedy:** With probability $\epsilon$, explore uniformly at random; with probability $1 - \epsilon$ act greedily according to $f_t$ (Sutton, 1996). Experimentally, we found $\epsilon = 0$ optimal on average, consistent with the results of Bietti et al. (2018).

**ε-decreasing:** Selects a random action with probabilities $\epsilon_i$, where $\epsilon_i = \epsilon_0 / t$, $\epsilon_0 \in [0, 1]$ and $t$ is the index of the current round. In our experiments we set $\epsilon_0 = 0.1$. (Sutton & Barto, 1998)

**Exponentiated Gradient ε-greedy:** Maintains a set of candidate values for ε-greedy exploration. At each iteration, it runs a sampling procedure to select a new $\epsilon$ from a finite set of candidates. The probabilities associated with the candidates are initialized uniformly and updated with the Exponentiated Gradient (EG) algorithm. Following Li et al. (2010b), we use the candidate set $\{\epsilon_i = 0.05 \times i + 0.01, i = 1, \ldots, 10\}$ for $\epsilon$.

**LinUCB:** Maintains confidence bounds for reward payoffs and selects actions with the highest confidence bound. It is impractical to run “as is” due to high-dimensional matrix inversions. We use diagonal approximation to the covariance when the dimensions exceeds 150. (Li et al., 2010a)

**τ-first:** Explore uniformly on the first $\tau$ fraction of the data; after that, act greedily.

**Cover:** Maintains a uniform distribution over a fixed number of policies. The policies are used to approximate a covering distribution over policies that are good for both exploration and exploitation (Agarwal et al., 2014).

**Cover Non-Uniform:** Similar to Cover, but reduces the level of exploration of Cover to be more competitive with the Greedy method. Cover-Nu doesn’t add extra exploration beyond the actions chose by the covering policies (Bietti et al., 2018).

In all cases, we select the best hyperparameters for each exploration algorithm following Bietti et al. (2018). These hyperparameters are: the choice of $\epsilon$ in ε-greedy, $\tau$ in τ-first, the number of bags, and the tolerance $\psi$ for Cover and Cover-NU. We set $\epsilon = 0.0$, $\tau = 0.02$, bag size = 16, and $\psi = 0.1$. 

A.4 Baseline Exploration Algorithms