Learning how contextual bandits learn

Anonymous Author(s)
Affiliation
Address
email

Abstract

In cognitive modeling, understanding how an agent leverages contextual information to learn about an adversarial environment and take what it considers good decisions is a fundamental investigation. By observing the agent's learning process, can we estimate how the agent is using this contextual information? One way of doing this is to approximate the agent's learning behavior by contextual bandit algorithms. The aim of this work is to provide model selection procedures that will pick the contextual bandit procedure that best fits the agent's learning process. We introduce a hold-out estimator and a penalized maximum likelihood estimator and show that both satisfy oracle inequalities. We give several examples of bandit algorithms for which the assumptions are satisfied, and assess our results on both synthetic and experimental learning data in a human categorization task. We also discuss why bandits with expert advice satisfy the same type of oracle inequalities and how they can be used to model metalearning in cognition.

1 Introduction

1.1 Cognitive models

Imagine an agent (human or animal) learning sequentially to make good decisions and having access at each time step to some contextual information. By looking at the agent's successive actions, can we estimate the agent's learning strategy, that is the way the agent used this contextual information to make its decisions? This problem belongs to the more general framework of cognitive modeling [12]. Cognitive models help to understand the mechanisms that occur while for instance learning, remembering or predicting tasks. They have been widely studied in the cognition literature [32, 15] and have a major impact on education for example. Usually in cognitive modeling [42, 16], maximum likelihood estimation (MLE) is applied and the best cognitive model is selected by cross-validation or an Akaike information criterion (AIC). One of the main challenges of cognitive modeling on learning data is that, since the agent remembers its past actions to learn, the data are not stationary and not independent. There are very few theoretical statistical works in this context: in [5], the properties of the MLE are studied for the `Exp3` model on learning data; in [6], a very general model selection procedure is presented that can be applied to non stationary data, but nothing in the setup of learning with contextual information. Our present goal is to provide model selection procedures that are valid for learning data in this contextual setup.

1.2 Contextual bandits

The purpose of a contextual bandit algorithm [28] is to find an optimal policy for selecting actions based on additional information (the context) given at each time step. In Machine Learning, contextual bandits have many applications [10] such as recommendation, patient follow-up in healthcare, etc. Here, we use them as learning models. Although not traditionally employed in cognition for modeling real behavioral data, contextual bandits are gaining popularity in the cognition literature [27, 41] and most of the cognitive psychology models of learning with contextual data such as Component Cue [20] or Alcove [26] can be expressed as contextual bandit algorithms since they treat the same problem: bandit feedback and choice based on past decisions and present context.

Submitted to 38th Conference on Neural Information Processing Systems (NeurIPS 2024). Do not distribute.

Let us formalize the statistical problem we treat. We observe a sequence of contexts and actions $(X_1, A_1, \dots, X_T, A_T)$ for an integer $T \geqslant 1$, where the contexts X_t belong to some finite space $\mathcal{X}$ and the actions A_t belong to a finite set $[K] = \{1, ..., K\}$ for some integer $K \geqslant 1$. Let $\mathcal{F}_0$ be the trivial sigma-algebra and for $t \geqslant 1$, let $\mathcal{F}_t = \sigma(X_1, A_1, \dots, X_t, A_t)$ be the sigma-algebra generated by observations until time t. Let $p^\star = (p^\star_t)_{t\in[T]}$ be the successive conditional probability distributions: $\forall x \in \mathcal{X}, \forall a \in [K], \quad p^\star_t(a, x) = \mathbb{P}(A_t = a|X_t = x, \mathcal{F}_{t-1})$. In reinforcement learning, this vector is called the *policy* of the agent. Recall that here, $p^\star$ is fixed, but unknown.
Our goal is to select the best model approximating $p^\star$ among a family of models $(\{p^m = (p^m_t)_{t\in[T]}\})_{m\in\mathcal{M}}$, where $\mathcal{M}$ is a countable set. Each p^m_t is a conditional distribution over $[K]$ given $(X_t, \mathcal{F}_{t-1})$ and a candidate at being $p^\star_t$.

1.3 Partition-based contextual bandits: an example of parametric models

The leading example of contextual bandit algorithm that we use here is *partition-based contextual bandits* [28, Chapter 18]. It consists in assuming that the agent partitions the context space $\mathcal{X}$ into disjoint cells C. This may typically happen if the agent is already familiar with the contexts and has already built a personal opinion on their meaning. The agent only has to learn the new task thanks to this fixed view of the space by updating elementary bandit algorithms in each cell C, that we denote $\texttt{CellBandit}(C)$, each time the context belongs to the corresponding cell. Our goal is to estimate the partitioning of the context space that the agent is using, *i.e.* understanding how the agent uses the contexts for the learning task. As an example, we illustrate numerically our approach on a categorization task, see Section 5 where contexts are objects to classify. By selecting the partition that best fits the learning data of a given individual, we have access to the similarity between objects as perceived by the learner.

To formalize *partition-based contextual bandits*, let $g_t = (g_{1,t}, \dots, g_{K,t}) \in [0, 1]^K$ be the vector of losses (or rewards) at time t, which models the feedback of the environment. We make no particular assumptions on the way losses are generated, except that g_t needs to be $\sigma(X_t, \mathcal{F}_{t-1})$-measurable. They may be adversarial or stochastic (see Section 4 for some examples). In the same way, the generation of the contexts X_t does not need to be specified: they can be independent of past actions or the result of the past actions. Then, each model $m \in \mathcal{M}$ corresponds to a partition $\mathcal{P}_m$ of $\mathcal{X}$ into D_m cells. The model m is parameterized by a vector $\theta^m = (\theta_C)_{C\in\mathcal{P}_m}$, where each $\texttt{CellBandit}(C)$ is using a procedure parameterized by a parameter θ_C – for instance, the learning rate in Exp3. The resulting candidate for $p^\star$ is therefore $p^m_{\theta^m} = (p^m_{\theta^m,t})_{t\in[T]}$. For a given cell $C \in \mathcal{P}_m$, $\texttt{CellBandit}(C)$ is updated each time $X_t \in C$, and therefore its decision at time t only depends on the contexts and actions happening at times in $F_t(C) = \{s \in [t], X_s \in C\}$, which is of cardinality $T^C_t = |F_t(C)|$. We write $a \mapsto \pi^{\theta_C}_{C,T^C_t}(a)$ the distribution over the set of actions $[K]$ at time t for the procedure $\texttt{CellBandit}(C)$ with parameter θ_C (see Algorithm 1). With this notation,

$$\forall t \in [T], \forall a \in [K], \quad p^m_{\theta^m,t}(a, X_t) = \mathbb{P}^m_{\theta^m}(A_t = a|X_t, \mathcal{F}_{t-1}) = \sum_{C\in\mathcal{P}_m} \pi^{\theta_C}_{C,T^C_t}(a)\mathbf{1}_{X_t\in C}. \quad (1)$$

Algorithm 1 Partition-based contextual bandit for model m [28]

Inputs: partition $\mathcal{P}_m$ of the context space $\mathcal{X}$,
parameters $\theta^m = (\theta_C)_{C\in\mathcal{P}_m} \in \Theta^m = \bigotimes_{C\in\mathcal{P}_m} \Theta_C$, with Θ_C compact parametric set.
Initialization: For all $C \in \mathcal{P}_m$, for all $a \in [K]$, $\pi^{\theta_C}_{C,1}(a) = 1/K$.
for $t = 1, 2, \dots$ **do**
Learner observes context $X_t \in \mathcal{X}$ and finds $C \in \mathcal{P}_m$ such that $X_t \in C$.
Learner plays $\texttt{CellBandit}(C)$ with parameter θ_C and samples action $A_t \sim \pi^{\theta_C}_{C,T^C_t}$.
Learner observes loss $g_{A_t,t}$ and updates the probability distribution $\pi^{\theta_C}_{C,T^C_t}$ in $\texttt{CellBandit}(C)$.

1.4 Contributions

We provide two model selection procedures for modeling learning with contextual information, based on the partial log-likelihood $\ell_T(p^m)$ of the observations $(X_1, A_1, \dots, X_T, A_T)$, defined by

$$\ell_T(p^m) = \sum_{t=1}^{T} \log\left(p_t^m(A_t, X_t)\right). \tag{2}$$

We prove oracle inequalities for the conditional Kullback-Leibler divergence $\mathrm{D_{KL}}$ between $p_t^\star(\cdot, X_t)$ and $p_t^m(\cdot, X_t)$:

$$\mathrm{D_{KL}}\left(p_t^\star(\cdot, X_t), p_t^m(\cdot, X_t)\right) = \mathbb{E}\left[\log \frac{p_t^\star(\cdot, X_t)}{p_t^m(\cdot, X_t)} \Big| X_t, \mathcal{F}_{t-1}\right].$$

In Section 2 we consider a finite family of general models $\{p^m = (p_t^m)_{t\in[T]}, m \in \mathcal{M}\}$ and show that a hold-out estimator satisfies an oracle inequality with an $\mathcal{O}\left((\log T + \log|\mathcal{M}|)/T\right)$ error bound, regardless of the nature of the models. In Section 3, we focus on the partition-based contextual bandit models defined in (1) with possibly infinite countable family of partitions and consider a log-likelihood criterion penalized by D_m times some logarithmic terms. Under some assumptions on the `CellBandit` algorithms that are used, we show an oracle inequality with an $\mathcal{O}\left(\log(T)^3/T\right)$ error bound. In Section 4, we prove that Stochastic Gradient Bandits and Exp3-IX are examples of `CellBandit` for which assumptions of Section 3 are satisfied. Section 5 is devoted to numerical illustrations on both synthetic and experimental learning data in a categorization task. In Section 6, we discuss how bandits with expert advice can be used to model metalearning [9], which refers to the processes by which an agent acquires knowledge about its own learning abilities, strategies, and preferences. In Appendix B, we give the details required to obtain model selection results for metalearning. The complete proofs of the theoretical results are given in Appendix C.

1.5 Related work

Our objective is not to provide a method that improves the regret [11]. Similarly, our work is not to be misunderstood with [17, 18, 37] in which authors develop model selection algorithms for contextual bandits that aim at finding the relation between context and action that best optimizes rewards. Our goal is to understand how an agent learns, not to tell it how to learn better. Thanks to the learning data of an agent, we select the contextual bandit algorithm that best fits the learning curve of the agent – without necessarily assuming that the agent understands the relation between context and actions. Hence we are not trying to find an optimal model, but the most realistic one w.r.t. learning data. To our knowledge, this theoretical statistical problem was studied for the first time in [5]. But in contrast with [5], which assumes `Exp3` to be true and studies MLE performances, we want to perform model selection with contexts.

From an Imitation Learning (IL) or Inverse Reinforcement Learning (IRL) point of view, this problem could be seen as a learner trying to reproduce the learning curve of an expert. Usually in IL [23], we observe an expert who has already mastered the task, so the input data of a classic IL algorithm are not learning data. In IRL [4], MLE might be used on data [38] but the IRL learner's goal is to infer the underlying reward function that best explains the expert's observed behavior thanks to multiple trajectories and then use this inferred reward function to guide its own decision-making. In our setting, the experimentalist already knows the reward function and the goal is to infer the agent's perception of the contexts, thanks to a single learning trajectory.

Our goal is close to [24], who estimate how a learner's behavior evolves over time and how it prioritises choices for applications to healthcare, except that [24] is in a Bayesian framework. Similarly, authors in [40] and [41] try to predict the behavior of participants in contextual multi-armed bandit tasks. The main difference is that they work in specific stochastic bandit settings with a Bayesian approach whereas we do model selection in a non-stationary and adversarial framework.

On a more technical level, hold-out estimators are often used in cognition for learning data [33, 25]. Hold-out procedures have been studied theoretically in the literature in a stationary and independent data framework [29, 3, 2]. Few results exist for time dependent data [36] and they are quite far from our setup. Here, the main issue is that the training set is not independent from the validation set, so more advanced tools such as V-fold cross-validation cannot be used.

Section 1.3 is very similar in design to the framework of [13] for selecting the best histogram for density estimation or more generally to non asymptotic model selection [29]. The main difference is that we are in a non stationary and non independent framework. Therefore, to prove the oracle inequality of Section 3,we use instead a recent result for penalized log-likelihood estimators which is valid in this framework [6].

2 Hold-out estimator

In this section, we assume that $\mathcal{M}$ is finite as it is often the case for hold-out estimators [29, Chapter 8], and $|\mathcal{M}| \geqslant 2$. Let $T > N \geqslant 1$ and select $\hat{m} \in \arg\max_{m \in \mathcal{M}} \sum_{t=N}^{T} \log p_t^m(A_t, X_t)$.

Theorem 1. *Assume that for all $m \in \mathcal{M}$ and for all $t \in \{N, ..., T\}$, p_t^m depends only on $(X_t, \mathcal{F}_{t-1})$. There exists a positive numerical constant $\diamondsuit$, such that for any $\kappa \in (0, 1)$,*

$$(1-\kappa)\mathbb{E}\left[\frac{1}{T-N+1}\sum_{s=N}^{T} \mathrm{D}_{\mathrm{KL}}\left(p_s^\star(\cdot, X_s), \frac{p_s^\star(\cdot, X_s) + p_s^{\hat{m}}(\cdot, X_s)}{2}\right) \middle| X_N, \mathcal{F}_{N-1}\right]$$
$$\leqslant (1+\kappa) \inf_{m \in \mathcal{M}} \mathbb{E}\left[\frac{1}{T-N+1}\sum_{s=N}^{T} \mathrm{D}_{\mathrm{KL}}\left(p_s^\star(\cdot, X_s), p_s^m(\cdot, X_s)\right) \middle| X_N, \mathcal{F}_{N-1}\right]$$
$$+ \frac{\diamondsuit}{\kappa} \frac{\log(T-N+1) + \log|\mathcal{M}|}{T-N+1}.$$

This result can hold for arbitrary p^m as long as it is adapted to $\mathcal{F}_t$, for $t \geqslant N$. In particular it allows, as usual for hold-out estimator, to use $p^m = p^m_{\tilde{\theta}^m}$, where $\tilde{\theta}^m \in \arg\max_{\theta^m \in \Theta^m} \sum_{t=1}^{N-1} \log p^m_{\theta^m, t}(A_t, X_t)$, whatever the parameterization of the model m – not necessarily partition-based. This result is the equivalent of Theorem 8.9 in [29] for this learning framework, adding only a multiplicative $\log T$ factor in the error bound. It justifies the use of hold-out procedures to model learning data in cognitive experiments such as [33, 25], using classical cognitive models as Alcove [26], Component-Cue [20] or Activity-based Credit Assignment (see [25] and the references therein).

Limitations. Due to the strongly dependent structure of the data, we perform a single split of the sample between training and testing data at $t = N$, unlike the classical hold-out. As usual, a careful trade-off has to be performed between N large enough to properly estimate each model and not too large, in order to reliably compare them. This also means that this approach is unsuited to situations where the learner learns differently at the start and at the end of the experiment, for instance by switching models once it has grasped how the task worked.

3 Penalized maximum likelihood estimator

In this section, we restrict ourselves to partition-based contextual bandit (see (1)). Following [5], we need to assume that the probabilities do not vanish.

Assumption 1. There exists $\varepsilon > 0$ and an integer $T_\varepsilon \geqslant 2$, such that, almost surely,

$$\forall t \leqslant T_\varepsilon,\ \forall x \in \mathcal{X},\ \forall a \in [K], \quad p_t^\star(a, x) \geqslant \varepsilon \tag{3}$$

and that for all $m \in \mathcal{M}$ and all $C \in \mathcal{P}_m$, the `CellBandit`(C) satisfies, for all parameter $\theta_C \in \Theta_C$

$$\forall t \leqslant T_\varepsilon,\ \forall a \in [K], \quad \pi^{\theta_C}_{C, T_t^C}(a) \geqslant \varepsilon. \tag{4}$$

Let pen : $\mathcal{M} \to \mathbb{R}_+$ be a penalty function. For each $m \in \mathcal{M}$, let $\hat{\theta}^m \in \arg\max_{\theta^m \in \Theta^m} \ell_{T_\varepsilon}(p^m_{\theta^m})$ be a MLE of model m, with ℓ defined as in (2), and select a model $\hat{m}$ that minimizes the penalized log-likelihood stopped at T_ε:

$$\widehat{m} \in \arg\min_{m \in \mathcal{M}} \left(-\frac{\ell_{T_\varepsilon}(p^m_{\hat{\theta}^m})}{T_\varepsilon} + \mathrm{pen}(m)\right). \tag{5}$$

To prove oracle inequalities, we also need a smoothness assumption on the parameterization of `CellBandit`(C) which can then be propagated to the p^m in Proposition 2.

Assumption 2. With the notation of Assumption 1, there exists $L_\varepsilon > 0$ such that, almost surely, for all $m \in \mathcal{M}$, all $C \in \mathcal{P}_m$,

$$\forall \delta_C, \theta_C \in \Theta_C,\ \forall t \leqslant T_\varepsilon, \quad \sup_{a \in [K]} \left|\log\left(\frac{\pi^{\delta_C}_{C, T_t^C}(a)}{\pi^{\theta_C}_{C, T_t^C}(a)}\right)\right| \leqslant L_\varepsilon \|\delta_C - \theta_C\|_2. \tag{6}$$

Proposition 2. *Assume that p^m is a partition-based contextual bandit as in* (1) *or Algorithm 1 and that there exists T_ε such that for all $C \in \mathcal{P}_m$,* `CellBandit`(C) *satisfies* (4) *and* (6)*. Then, almost surely, for all $\theta^m, \delta^m \in \Theta^m$, for all $t \leqslant T_\varepsilon$, for all $x \in \mathcal{X}$, for all $a \in [K]$,*

$$p^m_{\theta^m,t}(a,x) \geqslant \varepsilon \qquad \textit{and} \qquad \sup_{a\in[K]} \left| \log\left(\frac{p^m_{\delta^m,t}(a,x)}{p^m_{\theta^m,t}(a,x)}\right)\right| \leqslant L_\varepsilon \sup_{C\in\mathcal{P}_m} \|\delta_C - \theta_C\|_2.$$

Assume that the numbers of parameters of all `CellBandit` procedures are uniformly bounded, and let $d = \sup_{m\in\mathcal{M}} \sup_{C\in\mathcal{P}_m} \dim(\Theta_C)$. Since the models are smooth enough and the probabilities are lower bounded, by applying [6], one can prove the following result.

Theorem 3. *Let $\mathcal{M}$ be a countable set, and for each $m \in \mathcal{M}$, consider a partition-based contextual bandit model $\{p^m_{\theta^m}, \theta^m \in \Theta^m\}$ (see Algorithm 1 and* (1)*). Let R and r be such that all coordinates $\theta_{i,C}$'s of $\theta_C \in \Theta_C$, for $C \in \mathcal{P}_m$ and $m \in \mathcal{M}$, satisfy $r \leqslant \theta_{i,C} \leqslant R$ and let $A_\varepsilon = L_\varepsilon\sqrt{d}(R - r) + 2\log(\varepsilon^{-1})$. Let $\Sigma_\varepsilon = \log(A_\varepsilon)\sum_{m\in\mathcal{M}} e^{-D_m} < +\infty$. Under Assumptions 1 and 2, there exist positive numerical constants c and c' such that for all $\kappa \in (0,1]$, the following holds: if for all $m \in \mathcal{M}$,*

$$\textit{pen}(m) \geqslant \frac{c}{\kappa} A_\varepsilon^2 \log(\varepsilon^{-1})^{3/2} \log(T_\varepsilon A_\varepsilon)^2 \frac{D_m}{T_\varepsilon},$$

then,

$$\begin{aligned}\frac{1-\kappa}{T_\varepsilon}\sum_{t=1}^{T_\varepsilon} \mathbb{E}\left[\mathrm{D_{KL}}\left(p^\star_t(\cdot,X_t), p^{\hat m}_{\hat\theta^{\hat m},t}(\cdot,X_t)\right)\right] \\ \leqslant \inf_{m\in\mathcal{M}}\left((1+\kappa)\inf_{\theta^m\in\Theta^m}\frac{1}{T_\varepsilon}\sum_{t=1}^{T_\varepsilon}\mathbb{E}\left[\mathrm{D_{KL}}\left(p^\star_t(\cdot,X_t), p^m_{\theta^m,t}(\cdot,X_t)\right)\right] + 2\,\textit{pen}(m)\right) \\ + \frac{18c'}{\kappa} A_\varepsilon \Sigma_\varepsilon \log(\varepsilon^{-1})^{3/2}\log(T_\varepsilon A_\varepsilon)^2 \frac{\log(T_\varepsilon)}{T_\varepsilon}.\end{aligned}$$

This result is very similar to model selection "à la Birgé-Massart" [29, Section 7.4] with a trade-off between bias and variance represented by the penalty in D_m/T_ε, with additional logarithmic terms in $\log^2 T_\varepsilon$ in the penalty and in $\log^3 T_\varepsilon$ in the residual error. It is obtained by applying the recent and very general result of [6] which holds even for dependent non stationary data. However it is quite tedious to validate the assumptions of [6]. The partition-based contextual bandits are an example where this holds easily thanks to the partition which involves easier assumptions (namely (4) and (6)) to check on the `CellBandit` (see Section 4 for examples of `CellBandit` that satisfy them). Another example of contextual bandit where the assumptions of [6] are satisfied is given Section 6 in metalearning.

Limitations. Compared to the hold-out procedure, this approach does not require to split the sample. While this estimator can still work well when using all data, as shown in Section 5, the oracle inequality only holds when using the data from the time interval $[T_\varepsilon]$, which can be significantly less: in the models of Section 4, T_ε is of order $\sqrt{T}$. Moreover, this theorem does not cover every kind of cognitive models. Finally, while the penalty is in $c\log^2(T_\varepsilon)D_m/T_\varepsilon$, the constant c in this theoretical result is not known a priori and one needs to calibrate it by numerical simulations (see Section 5). We could use the hold-out procedure described in Section 2 to choose it, with similar issues, or other heuristics such as the dimension jump method or slope heuristics [7, 1].

4 Examples of `CellBandit`

In this section we provide examples of `CellBandit` satisfying (4) and (6). All the algorithms below are written for a cell C and a `CellBandit`(C) parameterized by $\theta_C \in \Theta_C$ compact subset of $\mathbb{R}^d$ such that $R \geqslant \sup_{\theta_C\in\Theta_C} \|\theta_C\|_\infty$.

4.1 Example 1: `Exp3-IX`

This algorithm is a generalization of `Exp3` and was introduced in [35]. Following [5], we write `Exp3-IX` with parameters decreasing as a square root of the sample size to ensure a good MLE estimation of the parameters. Note in addition that, for `Exp3` and its variants, it is well known that sublinear convergence of the regret occurs when the learning rate η and the exploration term γ are decreasing as a square root of the sample size. This renormalization ensures that the learner is able to learn at a good pace and at the same time be robust to changes in the environment.

Algorithm 2 `Exp3-IX`[35] as a `CellBandit`(C)

Inputs: T (Sample size), $\theta_C = (\eta, \gamma) \in \Theta_C$ (Parameter), K (Number of actions).
Initialization: $\pi^{\theta_C}_{C,1} = \left(\frac{1}{K}, \dots, \frac{1}{K}\right)$.
for $t \in F_T(C)$, the set of times where $X_s \in C$, **do**
  Draw an action $A_t \sim \pi^{\theta_C}_{C,T^C_t}$ and receive a loss $g_{A_t,t} \in [0,1]$.
  Update for all $a \in [K]$,

$$\pi^{\theta_C}_{C,T^C_t+1}(a) = \frac{\exp\left(-\frac{\eta}{\sqrt{T}} \sum_{s\in F_t(C)} \hat{g}^{\theta_C}_{a,s}\right)}{\sum_{b\in[K]} \exp\left(-\frac{\eta}{\sqrt{T}} \sum_{s\in F_t(C)} \hat{g}^{\theta_C}_{b,s}\right)} \quad \text{where } \hat{g}^{\theta_C}_{b,s} = \frac{g_{b,s}}{\gamma/\sqrt{T} + \pi^{\theta_C}_{C,T^C_s}(b)} \mathbf{1}_{A_s=b}$$

In this case, $\Theta_C \subset \mathbb{R}^2$. When $\gamma = 0$, we recover the classical `Exp3` algorithm, studied from the MLE point of view in [5]. Note that while $g_{A_t,t}$ is observed and known, the estimated loss $\hat{g}^{\theta}_{b,s}$ depends on the parameterization. The following result shows that one can choose `Exp3-IX` as a `CellBandit` in the partition-based contextual bandits to perform partition selection.

Proposition 4. *Let $\varepsilon \in (0, 1/K)$ and let $\Theta_C \subset [0, R]^2$ with $R > 0$. Then* `Exp3 − IX` *can be a* `CellBandit`(C) *with parameterization $\theta_C \in \Theta_C$ that satisfies* (4) *and* (6)*, as soon as*

$$T_\varepsilon = \left\lfloor \left(\frac{1}{K} - \varepsilon\right) \frac{\sqrt{T}}{R} \right\rfloor \wedge T \qquad \textit{and} \qquad L_\varepsilon = \frac{\sqrt{R^2/T + \varepsilon^2}}{\varepsilon^3 R} e^{1/\varepsilon^2}.$$

This shows that one can apply Theorem 3 with `Exp3-IX` as `CellBandit` as long as we stop using observations after $\sqrt{T}$ time steps. The dependence in ε in not very critical, since it has been proved at least for `Exp3` in [5], that in practice, we may take ε quite large (non-vanishing) with almost no impact on T_ε. This is a good thing since the theoretical dependency of L_ε in ε is quite pessimistic.

Limitations. This algorithm considers the horizon T fixed in order to renormalize the parameterization. From Proposition 4, it follows that Theorem 3 holds when only the first $\sqrt{T}$ observations are used in the MLE, but this in no way means that the estimator will perform poorly when based on all data. Taking $\sqrt{T}$ observations compounds on the usual issue that if the number of cells is large, only a small amount of data may be available for each cell, making estimation difficult.

4.2 Example 2: `Gradient Bandit`

`Gradient Bandit` is another possible algorithm. We still choose for similar reason a parameterization in $\eta/\sqrt{T}$, which echoes the Robbins-Monro conditions [39] even if [31] proved convergence in a stochastic bandit framework even for non renormalized parameters.

Algorithm 3 `Gradient Bandit` [31] as a `CellBandit`

Inputs: T (Sample size), $\theta_C \in [r, R]$ (Parameter), K (Number of actions).
Initialization: $\pi^{\theta_C}_{C,1} = \left(\frac{1}{K}, \dots, \frac{1}{K}\right)$.
for $t \in F_T(C)$ **do**
  Draw an action $A_t \sim \pi^{\theta_C}_{C,T^C_t}$ and receive a reward $g_{A_t,t} \in [0,1]$.
  Update for all $a \in [K]$,

$$\pi^{\theta_C}_{C,T^C_t+1}(a) = \frac{\exp\left(-\frac{\theta_C}{\sqrt{T}} \sum_{s\in F_t(C)} \hat{g}^{\theta_C}_{a,s}\right)}{\sum_{b\in[K]} \exp\left(-\frac{\theta_C}{\sqrt{T}} \sum_{s\in F_t(C)} \hat{g}^{\theta_C}_{b,s}\right)} \quad \text{where } \hat{g}^{\theta_C}_{b,s} = \left(\mathbf{1}_{A_s=b} - \pi^{\theta_C}_{C,T^C_s}(b)\right) g_{A_s,s}$$

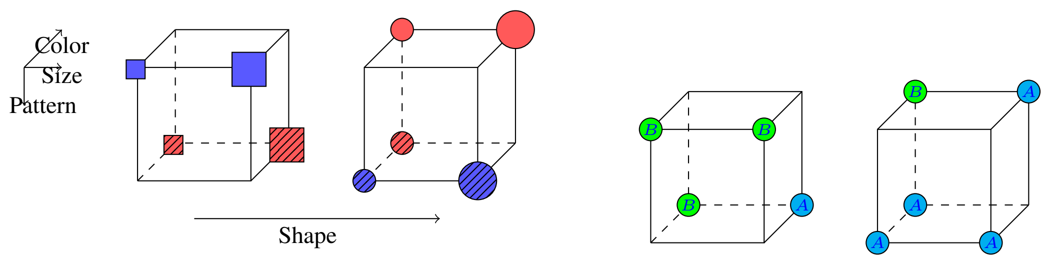
Proposition 5. *Let $\varepsilon \in (0,1)$ and let $\Theta_C \subset [0,R]^2$ with $R > 0$. Then,* `Gradient Bandit` *can be a* `CellBandit`(C) *with parameterization $\theta_C \in \Theta_C$ that satisfies* (4) *and* (6)*, as soon as*

$$T_\varepsilon := \left\lfloor \log\left(\sqrt{\frac{1}{K\varepsilon}}\right)\frac{\sqrt{T}}{R} \right\rfloor \wedge T \qquad \textit{and} \qquad L_\varepsilon = \frac{\sqrt{2}}{R\varepsilon}\frac{\log\left(\sqrt{\frac{1}{K\varepsilon}}\right)}{\sqrt{K\varepsilon}}.$$

This theoretical result has the same interpretation as before: the theoretical guarantees of Theorem 3 with `Gradient Bandit` as `CellBandit` hold when we stop using observations after $\sqrt{T}$ time steps. In practice, we can use the observations up to time T (see Section 5).

5 Numerical illustrations

We consider an experiment on the following categorization task: learners have to classify nine objects in two categories A and B in a sequential way. Figure 1 presents the objects and the classification rule the learners have to learn. It is a quite difficult task that has been experimented for instance in [34], where the learners needed about 300 trials to learn the classification rule.



(a) Representation in 4D space

(b) Category attribution

Figure 1: Experiment presentation: classic 5-4 category structure, widely used in cognition [30]. In 1a, the 9 objects to classify represented in a 4D space with respect to their attributes: Color, Size, Filling Pattern, and Shape. In 1b, by position in the 4D space, the category attribution (A or B).

The reward is fixed: 1 if the learner finds the good category and 0 in the other case. We focus on 6 different models (described in Table 1) that have a good cognitive interpretation.

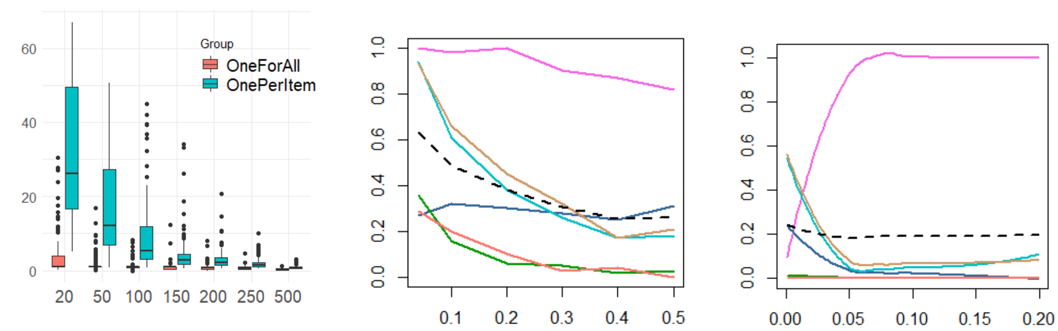
Table 1: Description of models and their learning abilities

Model	**Number of cells**	**Description of the cells**	**Learns categorization**
`OneForAll`	1	One giant cell	No
`ByShape`	2	One for circles, one for squares	Partly
`ByPattern`	2	One for striped items, one for plain items	Partly
`ByShapeExc`	4	Cells from `ByShape` model with exceptions isolated	Yes
`ByPatternExc`	4	Cells from `ByPattern` model with exceptions isolated	Yes
`OnePerItem`	9	One cell for each item	Yes

On synthetic data. We have not been able to run the simulation with `Exp3-IX`. Indeed, as also shown practically in [5] for the simple `Exp3` case, the probabilities $\pi^{\theta_C}_{C,T^C_t}$ can go to zero extremely fast. When the agent learns over an horizon $T = 500$, only $\sqrt{T} = 22$ observations would be usable and the estimations even of just the MLE is unreliable. So all the simulations were performed with

$T = 500$ and for `Gradient Bandit` as `CellBandit`, for all the 6 models described in Table 1. The way synthetic data are generated can be found in Appendix A.1.

Figure 2a shows that despite the conservative theoretical bound given in Proposition 5 with T_ε of order $\sqrt{T}$, `Gradient Bandit` provides good results when the MLE is applied to all T data points. The truncation at $\sqrt{T} \simeq 20$ required in the theoretical results does not seem necessary in practice, and actually looks suboptimal for `Gradient Bandit`. Figures 2b and 2c show that the hold-out is almost systematically outperformed by the penalized MLE. Both struggle to identify the significantly more complex model `OnePerItem`, preferring simpler alternatives.



(a) (b) (c)

Figure 2: Errors of the procedures as a function of the tuning parameters. In 2a, average of the $|\hat{\theta}_C - \theta_C|/\theta_C$ over all cells C in model `OneForAll` and `OnePerItem` for the data generated respectively by the same models, where $\hat{\theta}_C$ is the MLE with likelihood truncated at N (in abscissa). In 2b and 2c, percentage of mismatch between $\hat{m}$ and the simulated model over 100 simulations. The colors for each model are the ones given in Figure 3 whereas the average error on the models in the dash line. In 2b, for the hold-out estimator as a function of N/T. In 2c, for the penalized MLE with $\mathrm{pen}(m) = c\log(T)^2 D_m/T$, as a function of c.

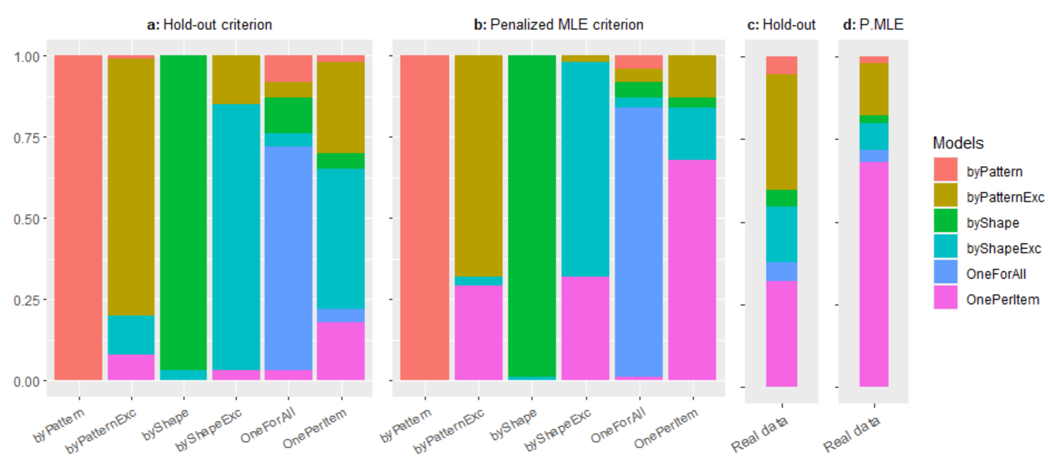


Figure 3: Distribution of the model choices. In **a**, hold-out with $N = 250$ over 100 simulations. In **b**, penalized MLE with $c = 0.012$ over 100 simulations. In **c**, hold-out on the data recorded in [34]–176 participants. In **d**, penalized MLE on the same experimental data.

Given these results on simulated data, we use $N = 250$ for the hold-out and $c = 0.012$ for the penalized MLE. The proportion of mismatches for each model are reported in Figure 3**a** for the hold-out and 3**b** for the penalized MLE. Both methods manage to recover the true model with less than 35% of mistakes, except for the model `OnePerItem`, for which only the penalized MLE is able to achieve a successful match more than 60% of the time. The models that are confused the most are the ones that are able to correctly learn the categorization, that is `ByPatternExc`, `ByShapeExc` and `OnePerItem`.

On real data. The data have been collected for [34][1] and we focus only on the learning data. We use only the 176 participants that needed at least $T = 100$ trials. In Figure 3**d**, we see that most of participants are attributed one of the 3 models able to learn. The most frequent is `OnePerItem` (about 70% for the penalized MLE) and this percentage is larger than the one obtained on simulation, probably meaning that a significant proportion of the participants do not see the division along the dimensions `Shape` or `Pattern`. It would be interesting for further study to see if this is linked to the presentation order of the objects, as it has been proved for Alcove and Component Cue in [34].

6 Metalearning

By looking at the experiment above, it is hard to believe that learners start directly with a model like `ByPatternExc`. It is more likely that they start with a model like `ByPattern` and realize that there are too many exceptions, so that they progressively end up with `ByPatternExp`. One way to model this progressive switch from one strategy to the other is to use bandits with expert advice. In this framework, there is a finite set E of randomized policies called experts, $(\xi_{j,t}(.))_{t\in[T]}$, probabilities over the set of actions $[k]$, that are modeling the different strategies the learner might have. No assumptions are made here on the way experts compute their randomized predictions: they might be the result of contextual bandits like `ByPattern` or more generally any kind of computations that depend on the learner's past choices. `Exp4` (see Algorithm 4) is an adaptation of `Exp3` to this case (see [28] for regret convergence and variants such as `Exp4.P` [8]).

Algorithm 4 Exp4 [11]

Inputs: T (Sample size), $\theta \in [r, R]$ (Parameter), K (Number of actions), E (Set of experts).
Initialization: $q^\theta_{E,1}$ uniform distribution over the experts E.
for $t = 1, 2, \ldots$ **do**
Receive experts advice $a \mapsto \xi_{j,t}(a)$ probability distribution over $[K]$ for all j.
Draw an action $A_t \sim \pi^\theta_{E,t}(.) = \sum_{j\in E} q^\theta_{E,t}(j)\xi_{j,t}(.)$ and receive a reward $g_{A_t,t} \in [0, 1]$.
Update for all $j \in E$,

$$q^\theta_{E,t+1}(j) = \frac{\exp\left(-\frac{\theta}{\sqrt{T}}\sum_{s=1}^t \hat{y}^\theta_{j,s}\right)}{\sum_{i\in E}\exp\left(-\frac{\theta}{\sqrt{T}}\sum_{s=1}^t \hat{y}^\theta_{i,s}\right)} \quad \text{with} \quad \hat{y}^\theta_{i,s} = \sum_{a\in[K]} \xi_{i,t}(a)\frac{g_{a,s}}{\pi^\theta_{E,s}(a)}\mathbf{1}_{A_s=a}$$

In this setting, a model m is defined by a finite set E_m representing the different experts/strategies the learner is learning from. Since there is only one parameter by model (namely $\theta \in [r, R]$), the penalty plays no role, nor the calibration of c. So there is no need for hold-out and one can prove that the model with the smallest log-likelihood on the first $T_\varepsilon \sim \sqrt{T}$ time steps satisfies an oracle inequality if $\mathcal{M}$ is finite, as well as $|F| := \max_{m\in\mathcal{M}} |E_m|$. Details are given in the Appendix B. This shows that one can select the set E_m of strategies which is the closest to reality among the sets of strategies that are put in competition.

Limitations. The only limitation with this approach is that we need at first to know the eventual parameters of each strategy. Again we could split the data in a hold-out fashion to make the injection of estimated parameters possible. However, it would be then nearly impossible to correctly estimate the parameters of strategies that are not used at the beginning of the learning. We refer to [9] for other methods in meta-learning for cognition.

References

[1] S. Arlot. Minimal penalties and the slope heuristics: a survey. *J. SFdS*, 160(3):158–168, 2019.

[2] S. Arlot and A. Celisse. A survey of cross-validation procedures for model selection. *Statistics Surveys*, 4:40–79, 2010.

[1] We refer the reader to [34] for precise description of the task as well as the ethics agreement.

[3] S. Arlot and M. Lerasle. Choice of V for V-fold cross-validation in least-squares density estimation. *J. Mach. Learn. Res.*, 17:Paper No. 208, 50, 2016.

[4] S. Arora and P. Doshi. A survey of inverse reinforcement learning: Challenges, methods and progress. *Artificial Intelligence*, 297:103500, 2021.

[5] J. Aubert, L. Lehéricy, and P. Reynaud-Bouret. On the convergence of the MLE as an estimator of the learning rate in the exp3 algorithm. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 1244–1275. PMLR, 7 2023.

[6] J. Aubert, L. Lehéricy, and P. Reynaud-Bouret. General oracle inequalities for a penalized log-likelihood criterion based on non-stationary data. Working paper or preprint, 5 2024.

[7] J.-P. Baudry, C. Maugis, and B. Michel. Slope heuristics: overview and implementation. *Statistics and Computing*, 22:455–470, 2012.

[8] A. Beygelzimer, J. Langford, L. Li, L. Reyzin, and R. Schapire. Contextual bandit algorithms with supervised learning guarantees. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 19–26. JMLR Workshop and Conference Proceedings, 2011.

[9] M. Binz, I. Dasgupta, A. K. Jagadish, M. Botvinick, J. X. Wang, and E. Schulz. Meta-learned models of cognition. *Behavioral and Brain Sciences*, pages 1–38, 2023.

[10] D. Bouneffouf and I. Rish. A survey on practical applications of multi-armed and contextual bandits. *arXiv preprint arXiv:1904.10040*, 2019.

[11] S. Bubeck and N. Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *CoRR*, abs/1204.5721, 2012.

[12] J. R. Busemeyer and A. Diederich. *Cognitive modeling*. Sage, 2010.

[13] G. Castellan. Density estimation via exponential model selection. *IEEE Trans. Inform. Theory*, 49(8):2052–2060, 2003.

[14] D. S. Clark. Short proof of a discrete gronwall inequality. *Discrete Applied Mathematics*, 16(3):279–281, 1987.

[15] A. G. Collins and A. Shenhav. Advances in modeling learning and decision-making in neuroscience. *Neuropsychopharmacology*, 47(1):104–118, 2022.

[16] N. D. Daw. Trial-by-trial data analysis using computational models. *Decision making, affect, and learning: Attention and performance XXIII*, 23(1), 2011.

[17] M. Dimakopoulou, Z. Zhou, S. Athey, and G. Imbens. Estimation considerations in contextual bandits. *arXiv preprint arXiv:1711.07077*, 2017.

[18] D. J. Foster, A. Krishnamurthy, and H. Luo. Model selection for contextual bandits. *Advances in Neural Information Processing Systems*, 32, 2019.

[19] B. Gao and L. Pavel. On the properties of the softmax function with application in game theory and reinforcement learning. *arXiv preprint arXiv:1704.00805*, 2017.

[20] M. A. Gluck and G. H. Bower. Evaluating an adaptive network model of human learning. *Journal of memory and Language*, 27(2):166–195, 1988.

[21] N. R. Hansen, P. Reynaud-Bouret, and V. Rivoirard. Lasso and probabilistic inequalities for multivariate point processes. *Bernoulli*, 21(1), 2 2015.

[22] C. Houdré and P. Reynaud-Bouret. Exponential inequalities for u-statistics of order two with constants. *Stochastic Inequalities and Applications. Progress in Probability*, 56, 01 2002.

[23] A. Hussein, M. M. Gaber, E. Elyan, and C. Jayne. Imitation learning: A survey of learning methods. *ACM Computing Surveys (CSUR)*, 50(2):1–35, 2017.

[24] A. Hüyük, D. Jarrett, and M. van der Schaar. Inverse contextual bandits: Learning how behavior evolves over time. In *International Conference on Machine Learning*, pages 9506–9524. PMLR, 2022.

[25] A. James, P. Reynaud-Bouret, G. Mezzadri, F. Sargolini, I. Bethus, and A. Muzy. Strategy inference during learning via cognitive activity-based credit assignment models. *Scientific reports*, 13(1):9408, 2023.

[26] J. K. Kruschke. Alcove: An exemplar-based connectionist model of category learning. In *Connectionist Psychology*, pages 107–138. Psychology Press, 2020.

[27] A. S. Lan and R. G. Baraniuk. A contextual bandits framework for personalized learning action selection. In *EDM*, pages 424–429, 2016.

[28] T. Lattimore and C. Szepesvári. *Bandit Algorithms*. Cambridge University Press, 2020.

[29] P. Massart. *Concentration inequalities and model selection: Ecole d'Eté de Probabilités de Saint-Flour XXXIII-2003*. Springer, 2007.

[30] D. L. Medin and M. M. Schaffer. Context theory of classification learning. *Psychological review*, 85(3):207, 1978.

[31] J. Mei, Z. Zhong, B. Dai, A. Agarwal, C. Szepesvari, and D. Schuurmans. Stochastic gradient succeeds for bandits. In *International Conference on Machine Learning*, pages 24325–24360. PMLR, 2023.

[32] G. Mezzadri. *Statistical inference for categorization models and presentation order*. Phd thesis, Université Côte d'Azur, LJAD, France, 2020.

[33] G. Mezzadri, T. Laloë, F. Mathy, and P. Reynaud-Bouret. Hold-out strategy for selecting learning models: application to categorization subjected to presentation orders. *Journal of Mathematical Psychology*, 109:102691, 2022.

[34] G. Mezzadri, P. Reynaud-Bouret, T. Laloë, and F. Mathy. Investigating interactions between types of order in categorization. *Scientific Reports*, 12(1):21625, 2022.

[35] G. Neu. Explore no more: Improved high-probability regret bounds for non-stochastic bandits. *Advances in Neural Information Processing Systems*, 28, 2015.

[36] J. Opsomer, Y. Wang, and Y. Yang. Nonparametric Regression with Correlated Errors. *Statistical Science*, 16(2):134 – 153, 2001.

[37] A. Pacchiano, M. Phan, Y. Abbasi Yadkori, A. Rao, J. Zimmert, T. Lattimore, and C. Szepesvari. Model selection in contextual stochastic bandit problems. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 10328–10337. Curran Associates, Inc., 2020.

[38] G. Ramponi, G. Drappo, and M. Restelli. Inverse reinforcement learning from a gradient-based learner. *Advances in Neural Information Processing Systems*, 33:2458–2468, 2020.

[39] H. Robbins and S. Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951.

[40] E. Schulz, E. Konstantinidis, and M. Speekenbrink. Learning and decisions in contextual multi-armed bandit tasks. In *CogSci*, 2015.

[41] E. Schulz, E. Konstantinidis, and M. Speekenbrink. Putting bandits into context: How function learning supports decision making. *Journal of experimental psychology: learning, memory, and cognition*, 44(6):927, 2018.

[42] R. C. Wilson and A. G. Collins. Ten simple rules for the computational modeling of behavioral data. *Elife*, 8:e49547, 2019.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We describe our contributions in the abstract and introduction. They are mainly theoretical in nature, with oracle inequalities for two model selection procedures (hold-out and penalized maximum likelihood) in Sections 2 and 3. We also show that the assumptions of these results are satisfied for some common reinforcement learning models in Section 4 and 6 and finally illustrate the procedure on synthetic and real data in Section 5.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: After each theoretical result, we wrote a paragraph named limitations to discuss the issues we thought of with respect to each result. Note that since our focus is on the theoretical properties of the methods, with the experimental sections serving as illustration of these properties, we do not care about the computational complexity of the algorithms considered and do not discuss them beyond their execution time discussed in Appendix A.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.

- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: We stated precisely all our theoretical results with assumptions that are clearly referenced. Only the last section about metalearning is written slightly more informally due to lack of space but formally written in the supplementary material. We give intuition about how each assumption is used. The proofs are given in the supplementary material.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The simulation of data according to each model is precisely explained in Appendix A, and the code to produce it is given. The codes to compute both estimators (hold-out and penalized MLE) are given and commented, also in the supplementary material. The real data are taken from a published paper [34] and we asked the authors of [34] to provide us with these data. We don't think it is possible to make these data public because the ethics agreement that has been signed by the authors of [34], prior to data collection, might not include this possibility. However, this application to real data is mainly to prove that this can be done in practice. Since there is not a truth to be compared to in these data and since even cross validation is hard to perform, this real dataset cannot be used as a benchmark to compare methods anyway.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.

- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: All codes to generate synthetic data and to perform penalized MLE and hold-out are provided, so that all the numerical part on the calibration of both methods can be faithfully reproduced.Only the real dataset, as explained before, cannot be given for reproduction (Figure 3).

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.

- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: The whole purpose of our numerical study in Section 5 is to explain the choice of hyperparameters (such as the splitting in the hold-out of the calibration of the constant c in the penalty).

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We have run our simulations on 600 independent simulated learners and we show with a boxplot (Figure 2a) and mismatch proportion graphs (Figure 2b, 2c and 3) the proportion of erroneous selections. This cannot be done on real data, since each real participant to the categorization task is unique.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.

- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: It is not central in our analysis so it is just mentionned in the supplementary material in Appendix A.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Up to the real data that are used in this paper, there is absolutely nothing that would be in violation to the Code of Ethics. For the real data that are used, they are human categorization data. They have been recorded for another publication and just transmitted to us. The experimental procedure was approved by the local ethics committee of the authors. We do not want to share these data publicly since we do not want to breach the Privacy rule of the Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: This work is theoretical. The methods that are validated theoretically here have already been in use in practice for a long time (see for instance the rules to follow for cognitive modeling in [42]) and so the expected societal impact of the present work is negligible.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.

- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: We do not think this applies to our research.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We clearly stated that the real data come from [34]. The code has been developed by us solely, using classical packages in `R` that are clearly mentioned in the code and supplementary material.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.

- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: We do not provide new packages associated to our results.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [No]

    Justification: We did not collect data for the present article but used data collected for [34], a work that is already published. In this article, all the details are given and we do not think it makes sense to reproduce it here for our illustration. We only kept the main description of the task so that the readers can understand what was done.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [No]

Justification: The data collection done for [34] had the approval of the local ethic committee as mentioned in their article. Here we do not feel necessary to reproduce this here but rather point towards [34] for additional information about the task and its ethic agreement.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

A Code and data description

A.1 Details on numerical illustrations

In this section, we give details on the numerical illustrations of Section 5. The images were obtained using the `ggplot2` package of `R`. Two types of analyses were conducted, on synthetic data and on real data.

On synthetic data. The simulations of the synthetic data helped us calibrate the tuning parameters choices for the hold-out and the penalized log-likelihood procedure. In Section 2, the parameter N must be calibrated for choosing the correct training data sample size. In Section 3, as said in the *Limitations*, since the constant c in the penalty term is not known a priori, it must be calibrated as well. To do this, we follow the guidelines of [42]. The procedure is as follows.

1) Sample size: $T = 500$. It is of the same order of magnitude as real data.

2) Objects generation: periodic sequence of the nine objects repeated through the T trials. We generate a sequence of objects following the same structure as in [34]. Due to the periodic pattern, each object is therefore seen roughly the same number of times for all time t.

3) Actions generation: for each model in Table 1, we generated 100 sequences of actions called synthetic agents with respect to the procedure given in Algorithm 1 with `Gradient Bandit` as `CellBandit`. The parameters θ_C we used were the same for each model and the same for each cell, equal to $0.03 \times \sqrt{T}$, except for the `OnePerItem` model where we changed slightly the values of the parameter in each cell to make the model identifiable. For $m =$`OnePerItem`, we took $\theta^m = ((0.03/10 + k \times 0.007) \times \sqrt{T})_{k \in \{0,\ldots,8\}}$ following the same order of presentation of the sequence of objects defined earlier.

4) Parameters estimation: we then fitted each of the six models on all the synthetic agents generated data, and we estimated the associated parameters using (MLE) and the package `DEoptim` in `R` with range $(0, 1)$ for the parameters $\theta_C/\sqrt{T}$ and with the default parameters and a maxiter value equal to 20. We then computed the log-likelihood associated to the estimated parameters. We did this for the likelihood stopped at time $N \in \{25, 50, 100, 150, 200, 250, T\}$.

   - With such data, we were able to plot Figure 2a and Figure 2b with the hold-out criterion defined in Section 2. In Figure 2a, we computed the average error made in each cell by the model fitting of the same model that generated the data. For the Figure 2b, we simply counted the number of times each model verified the hold-out criterion for all the synthetic agent and for each model that generated the data.

   - With the log-likelihood stopped at time T for the estimated parameters, we were able to plot Figure 2c according to the penalized log-likelihood criterion defined in (5). In the same way we counted the number of times each model satisfied the penalized log-likelihood criterion for all the synthetic agent and for each model that generated the data.

5) Choice of the parameters N and c for the real data: Given the results of Figure 2b and Figure 2c, we chose to use N to be equal to half of the data length and $c = 0.012$ to account for a reasonable error for model `OnePerItem`, even if in average $c = 0.04$ gives better results. With this data, we were able plot the two first chart of Figure 3.

On real data. For the real experimental data, here is the process we followed.

1) Sample size: dependent on each agent, the average data sample size is 300.

2) Objects and Actions: we collected for each agent their objects sequence and associated choices.

3) For each agent, we fitted the 6 models and estimated the parameters associated to each model. To perform hold-out and penalized log-likelihood model selection, we used the parameters N and c chosen thanks to the synthetic data. With this data, we were able to plot Figure 3.

A.2 About the code and the data

In this section, we give explanations about the code and data (e.g. computation time, link between code and data). All the data, code and images used are provided in the zip file associated to submission, called `ContextualBanditsCode`. We run all the simulations in `R` and used the following packages: `DEoptim`, `crayon`, `magrittr`, `dplyr`, `tidyr`, `ggplot2`, `gridExtra`.

For the sample size we chose, all the simulations can run on a PC in a reasonable time of execution (detailed hereafter). Overall, computing the different data and running the code took approximately 6 hours excluding the time needed for the real data. The biggest file is 373 kilobytes. The PC we used was a Gigabyte - AORUS 15G XC, with processor: Intel(R) Core(TM) i7-10870H CPU 2.20GHz, 2208 MHz, 8 cores, 16 logical processors.

On real data. As mentioned earlier, we could not provide the experimental data used in [34], since they have already been published in another paper and we do not want to break the ethic agreement. We can only provide the results and estimated data resulting from the experimental data. Note however that the procedures to obtain the following RData files are the same as for the synthetic data which we detail later. The three RData files on the real data are `realdatamle`, `realdata_holdout_trainingset`, `realdata_holdout_testingset`.

- `realdatamle` is a list of estimators and associated log-likelihood for each model and each agent.
- `realdata_holdout_trainingset` is a list of estimators and associated log-likelihood on the first half of the sample for each model and each agent.
- `realdata_holdout_testingset` is a list of log-likelihood on the testing part of the sample for each agent and each model with parameters estimated in `realdata_holdout_trainingset`.

On synthetic data. All the synthetic data obtained in the other files can be computed by running the code `ContextualbanditsCodebis`. The code is commented and starts with a list of functions which are necessary to run the different procedures. In the code, we explain how the different procedures lead to the following list of files. We have commented with # the parts of the code that would modify the files so that running the code now would give the same images as the ones used in the article. If one wants to generate new data, one should uncomment these lines of code. However, we advise the reader that some of the procedures take a certain time, and would recommend not to do so. We detail hereafter the content of the different csv and RData files and the time it took to run them.

- To begin with, we generate a csv file called `databis_500.csv` of 500 trials and associated list of objects in the file `synthetic_data`.
- In the same `synthetic_data` file we create the different model files and within each of them generate 100 csv files of actions, rewards, and objects according to the procedure described in A.1. This procedure takes around 5 minutes. Then, we begin to compute the MLE for each of the synthetic data csv file.
- `Datalikelihood100agents6modeletabis500horizon` is a nested list of estimators, associated log-likelihood stopped at time T for each model fitted to the data of all the synthetic agents. Computing these data took approximately 2 hours .
- `holdoutbis100agents6models_horizon_20` is the same nested list of estimators but computed on a log-likelihood stopped at time $N = 20$. Computing these data took approximately 10 minutes .
- `holdoutbis100agents6models_horizon_50` is the same nested list of estimators but computed on a log-likelihood stopped at time $N = 50$. Computing these data took approximately 20 minutes .
- `holdoutbis100agents6models_horizon_100` is the same nested list of estimators but computed on a log-likelihood stopped at time $N = 100$. Computing these data took approximately 30 minutes .

- `holdoutbis100agents6models_horizon_150` is the same nested list of estimators but computed on a log-likelihood stopped at time $N = 150$. Computing these data took approximately 40 minutes .
- `holdoutbis100agents6models_horizon_200` is the same nested list of estimators but computed on a log-likelihood stopped at time $N = 200$. Computing these data took approximately 50 minutes .
- `holdoutbis100agents6models_horizon_250` is the same nested list of estimators but computed on a log-likelihood stopped at time $N = 250$. Computing these data took approximately 1 hour .
- `alldataholdoutbis` is a nested list of errors on estimation for the training data and log-likelihood function on the testing data for all synthetic agents, all models and all training data sample size $N \in \{20, 50, 100, 150, 200, 250\}$. Computing these data took approximately 10 minutes .

B Assumptions for metalearning

Since we work in a more general setting and not simply with contexts, we assume that we observe a process $(A_t)_{1\leqslant t\leqslant T}$ adapted to a general filtration $(\mathcal{F}_t)_{1\leqslant t\leqslant T}$ where for all $t \in [T]$, $A_t \in [K]$. In particular, for all $t \in [T]$, $\mathcal{F}_t$ is generated by the past actions $(A_1, \dots, A_t)$ and any other additional variable which might be observed or not – such as a context at time $t + 1$ for instance. We write, for all $a \in [K]$, and all $t \in [T]$

$$p_t^\star(a) = \mathbb{P}(A_t = a|\mathcal{F}_{t-1})$$

the true conditional distribution we wish to estimate.

Additionally, we consider the family of models $\{(\pi^{\theta^m}_{E_m,t})_{t\in T}, m \in \mathcal{M}\}$ where $\mathcal{M}$ is a finite set, $\theta^m \in [r, R]$, and for all $m \in \mathcal{M}$, $(\pi^{\theta^m}_{E_m,t})_{t\in[T]}$ is the sequence of mixtures of probability distributions over actions defined recursively in Algorithm 4 for the finite set E_m. Each model m is thus defined by a set of experts $(\xi_{j,t}(.))_{j\in E_m,t\in[T]}$ where for all $m \in E_m$, $t \in [T]$, $\xi_{j,t}$ can be any probability distribution over arms $[K]$ as long as it is measurable with respect to $\mathcal{F}_{t-1}$.

Let $|F| := \max_{m\in\mathcal{M}} |E_m|$. The goal is to select the set E_m of policies – that we see as learning strategies – with which the agent learns to learn. This approach is again based on partial log-likelihood $\ell_T(\pi^{\theta^m}_{E_m})$ of the observations $(A_1, \dots, A_T)$ defined by

$$\ell_T(\pi^{\theta^m}_{E_m}) = \sum_{t=1}^{T} \log\left(\pi^{\theta^m}_{E_m}(A_t)\right). \tag{7}$$

To achieve a model selection result, we need the following assumption on the policies given by the experts.

Assumption 3. There exists $\rho > 0$, such that almost surely, for all $m \in \mathcal{M}$, for all $t \in [T]$ and all $i \in [K]$, $\sum_{j\in E_m} \xi_{j,t}(i) \geqslant \rho$.

Then, with Assumption 3, we can deduce a result similar to Propositions 4 and 5 because of the very structure of Algorithm 4 which mimics `Exp3`.

Proposition 6. *Assume Assumption 3 holds. Let ρ be the associated constant. Let $\varepsilon \in (0, \rho/|F|)$, and let*

$$T_\varepsilon = \left\lfloor \left(\frac{1}{|F|} - \frac{\varepsilon}{\rho}\right)\frac{\sqrt{T}}{R} \right\rfloor \wedge T \qquad \textit{and} \qquad L_\varepsilon = \frac{1}{R\varepsilon^2}\exp\left(\frac{1}{\varepsilon^2}\right).$$

Then, for all $t \in [T_\varepsilon]$, for all $m \in \mathcal{M}$, $\theta^m, \delta^m \in [r, R]$, for all $k \in [K]$,

$$\pi^{\theta^m}_{E_m,t}(k) \geqslant \varepsilon \qquad \textit{and} \qquad \sup_{k\in[K]}\left|\log\left(\frac{\pi^{\theta^m}_{E_m,t}(k)}{\pi^{\delta^m}_{E_m,t}(k)}\right)\right| \leqslant L_\varepsilon|\theta^m - \delta^m|.$$

Finally, we still assume that the true distribution is bounded away from 0 (as in (3)).

Assumption 4. Assume that Assumption 3 holds. Let ε and T_ε be the constants of Proposition 6. Assume that

$$\forall t \leqslant T_\varepsilon, \forall a \in [K], p_t^\star(a) \geqslant \varepsilon.$$

Assumptions 3 and 4 allow us to verify Assumptions 1 and 2 of [6]. As for Section 3, it is thus possible to put into competition different sets of experts. Let $A_\varepsilon = L_\varepsilon(R-r) + 2\log(\varepsilon^{-1})$. Since all the models have the same dimension, there is no penalty term to account for. So the term Σ_ε in Theorem 3 becomes $\log(A_\varepsilon)|\mathcal{M}|e^{-1}$. The result of [6, Corollary 2] states that there exist constants c, c' such that, for all $\kappa \in (0,1]$,

$$\begin{aligned}\frac{1-\kappa}{T_\varepsilon}\sum_{t=1}^{T_\varepsilon}\mathbb{E}\left[\mathrm{D_{KL}}\left(p_t^\star, \pi_{E_{\hat{m}},t}^{\hat{\theta}^{\hat{m}}}\right)\right] \leqslant \inf_{m\in\mathcal{M}}\left((1+\kappa)\inf_{\theta\in\Theta^{D_m}}\frac{1}{T_\varepsilon}\sum_{t=1}^{T_\varepsilon}\mathbb{E}\left[\mathrm{D_{KL}}\left(p_t^\star, \pi_{E_m,t}^{\theta^m}\right)\right]\right)\\ + \frac{c}{\kappa}A_\varepsilon^2\log(\varepsilon^{-1})^{3/2}\log(T_\varepsilon A_\varepsilon)^2\frac{1}{T_\varepsilon}\\ + \frac{18e^{-1}c'}{\kappa}A_\varepsilon\log(A_\varepsilon)|\mathcal{M}|\log(\varepsilon^{-1})^{3/2}\log(T_\varepsilon A_\varepsilon)^2\frac{\log(T_\varepsilon)}{T_\varepsilon}.\end{aligned}$$

C Proofs

C.1 Proof of Section 2

Proof of Theorem 1. For any $m \in \mathcal{M}$, and $k \in [K]$, let

$$\begin{cases} g_m(k, X_t) = -\dfrac{1}{2}\log\left(\dfrac{p_t^m(k,X_t)}{p_t^\star(k,X_t)}\right) \\ f_m(k,X_t) = -\log\left(\dfrac{p_t^\star(k,X_t) + p_t^m(k,X_t)}{2p_t^\star(k,X_t)}\right) \end{cases}.$$

For any function $h : [K]\times\mathcal{X} \to \mathbb{R}$, let

$$\begin{cases} \tilde{\nu}_T(h) = \dfrac{1}{T-N+1}\displaystyle\sum_{t=N}^{T}\sum_{k=1}^{K} h(k,X_t)\left(\mathbf{1}_{A_t=k} - p_t^\star(k,X_t)\right), \\ \tilde{P}_T(h) = \dfrac{1}{T-N+1}\displaystyle\sum_{t=N}^{T}\sum_{k=1}^{K} h(k,X_t)\mathbf{1}_{A_t=k}, \\ \tilde{C}_T(h) = \dfrac{1}{T-N+1}\displaystyle\sum_{t=N}^{T}\sum_{k=1}^{K} h(k,X_t)p_t^\star(k,X_t). \end{cases}$$

From the definition of $\hat{m}$,

$$\tilde{P}_T\left(g_{\hat{m}}\right) \leqslant \tilde{P}_T\left(g_m\right).$$

Since, $f_{\hat{m}}(k,X_t) \leqslant g_{\hat{m}}(k,X_t)$ by concavity of log, it holds that

$$\tilde{\nu}_T\left(f_{\hat{m}}\right) + \tilde{C}_T\left(f_{\hat{m}}\right) = \tilde{P}_T\left(f_{\hat{m}}\right) \leqslant \tilde{P}_T\left(g_m\right) = \tilde{\nu}_T\left(g_m\right) + \tilde{C}_T\left(g_m\right).$$

That is

$$\tilde{\nu}_T\left(f_{\hat{m}} - f_m\right) + \tilde{C}_T\left(f_{\hat{m}}\right) \leqslant \tilde{\nu}_T\left(g_m - f_m\right) + \tilde{C}_T\left(g_m\right).$$

Let $U_m = \tilde{\nu}_T\left(g_m - f_m\right)$, then

$$\tilde{C}_T\left(f_{\hat{m}}\right) \leqslant \tilde{C}_T\left(g_m\right) - \tilde{\nu}_T\left(f_{\hat{m}} - f_m\right) + U_m. \tag{8}$$

Note that U_m is centered. For $m' \in \mathcal{M}$, let $M_N^{m'} = 0$, and for $t \geqslant N+1$, let

$$M_t^{m'} = -\sum_{s=N}^{t-1}\sum_{k=1}^{K}\left(f_{m'}(k,X_s) - f_m(k,X_s)\right)\left(\mathbf{1}_{A_s=k} - p_s^\star(k,X_s)\right).$$

For all $t \geqslant N$, let $\mathcal{H}_t = \sigma(X_t, \mathcal{F}_{t-1})$. Then, $(M_t^{m'})_{t \geqslant N}$ is an $(\mathcal{H}_t)_{t \geqslant N}$-martingale. For $\ell \geqslant 2$, let $B_N^\ell = 0$, and for $t \geqslant N+1$, let

$$B_t^\ell := \sum_{s=N}^{t-1} \mathbb{E}\left[\left(M_{s+1}^{m'} - M_s^{m'} \right)^\ell \,\middle|\, \mathcal{H}_s \right].$$

For $t \in \{N, \dots, T-1\}$, note that

$$|M_{t+1}^{m'} - M_t^{m'}| \leqslant 2 \sum_{k=1}^{K} |f_{m'}(k, X_t) - f_m(k, X_t)| \, \frac{\mathbf{1}_{A_t = k} + p_t^\star(k, X_t)}{2},$$

so that, by convexity of $x \mapsto x^\ell$ on $[0, +\infty)$,

$$|M_{t+1}^{m'} - M_t^{m'}|^\ell \leqslant 2^\ell \sum_{k=1}^{K} |f_{m'}(k, X_t) - f_m(k, X_t)|^\ell \, \frac{\mathbf{1}_{A_t = k} + p_t^\star(k, X_t)}{2}$$

Thus,

$$\begin{aligned} B_t^\ell &= \sum_{s=N}^{t-1} \mathbb{E}\left[\left(M_{s+1}^{m'} - M_s^{m'} \right)^\ell \,\middle|\, \mathcal{H}_s \right] \\ &\leqslant 2^\ell \sum_{s=N}^{t-1} \sum_{k=1}^{K} |f_{m'}(k, X_s) - f_m(k, X_s)|^\ell \, p_s^\star(k, X_s) \\ &= 2^\ell \sum_{s=N}^{t-1} \sum_{k=1}^{K} \left| \log\left(\frac{p_s^\star(k, X_s) + p_s^{m'}(k, X_s)}{p_s^\star(k, X_s) + p_s^m(k, X_s)} \right) \right|^\ell p_s^\star(k, X_s). \end{aligned} \tag{9}$$

We now need the following Lemma to continue.

Lemma 7. *[29, Lemma 7.26] For all $\ell \geqslant 2$ and all $x > 0$,*

$$\frac{|\log(x)|^\ell}{\ell!} \leqslant \frac{9}{64} \left(x - \frac{1}{x} \right)^2.$$

Proof. The complete Lemma and proof of the Lemma can be found in [29]. $\square$

Applying Lemma 7 to $x = \sqrt{\dfrac{p_s^\star(k, X_s) + p_s^{m'}(k, X_s)}{p_s^\star(k, X_s) + p_s^m(k, X_s)}}$ leads to, for all $k \in [K]$,

$$\left| \log\left(\frac{p_s^\star(k, X_s) + p_s^{m'}(k, X_s)}{p_s^\star(k, X_s) + p_s^m(k, X_s)} \right) \right|^\ell \\ \leqslant \frac{9}{64} 2^\ell \ell! \frac{\left(p_s^m(k, X_s) - p_s^{m'}(k, X_s) \right)^2}{\left(p_s^m(k, X_s) + p_s^\star(k, X_s) \right) \left(p_s^\star(k, X_s) + p_s^{m'}(k, X_s) \right)}.$$

Plugging this in Equation (9) leads to

$$|B_t^\ell| \leqslant \frac{9}{4} 2^{2(\ell-2)} \ell! \sum_{s=N}^{t-1} \sum_{k=1}^{K} \frac{\left(p_s^m(k, X_s) - p_s^{m'}(k, X_s) \right)^2 p_s^\star(k, X_s)}{\left(p_s^m(k, X_s) + p_s^\star(k, X_s) \right) \left(p_s^\star(k, X_s) + p_s^{m'}(k, X_s) \right)}.$$

For all $x, y, z \geqslant 0$,

$$\left(\sqrt{x} + \sqrt{y} \right)^2 z \leqslant (z + y)(z + x),$$

therefore, with $z = p_s^\star(k, X_s)$, $x = p_s^m(k, X_s)$ and $y = p_s^{m'}(k, X_s)$,

$$|B_t^\ell| \leqslant \frac{9}{4} 4^{\ell-2} \ell! \sum_{s=N}^{t-1} \sum_{k=1}^{K} \left(\sqrt{p_s^m(k, X_s)} - \sqrt{p_s^{m'}(k, X_s)} \right)^2 \leqslant \frac{1}{2} 4^{\ell-2} \ell! V_t^{m'}, \tag{10}$$

where

$$V_t^{m'} := \frac{9}{2}\sum_{s=N}^{t-1}\sum_{k=1}^{K}\left(\sqrt{p_s^m(k,X_s)}-\sqrt{p_s^{m'}(k,X_s)}\right)^2$$
$$= 9\sum_{s=N}^{t-1} H\left(p_s^m(\cdot,X_s),p_s^{m'}(\cdot,X_s)\right)^2 \tag{11}$$

where H is the Hellinger distance between the two probability distributions $p_s^m(\cdot,X_s)$ and $p_s^{m'}(\cdot,X_s)$. Lemma 3.3 of [22] gives that for all $\lambda>0$,

$$\mathcal{E}_t = \exp\left(\lambda M_t^{m'} - \sum_{\ell\geqslant 2}\frac{\lambda^\ell}{\ell!}B_t^\ell\right)$$

is a supermartingale and that in particular $\mathbb{E}(\mathcal{E}_{T+1})\leqslant 1$. By Equation (10), for $\lambda\in(0,1/4)$,

$$\sum_{\ell\geqslant 2}\frac{\lambda^\ell}{\ell!}B_t^\ell \leqslant \frac{\lambda^2}{2}\sum_{\ell\geqslant 2}(4\lambda)^{\ell-2}V_t^{m'} = \frac{\lambda^2}{2(1-4\lambda)}V_t^{m'}.$$

Let $\Psi(\lambda)=\frac{\lambda^2}{2(1-4\lambda)}$ for $\lambda\in(0,1/4)$. Then,

$$\mathbb{E}\left[e^{\lambda M_{T+1}^{m'}-\Psi(\lambda)V_{T+1}^{m'}}\,\middle|\,\mathcal{H}_N\right]\leqslant 1.$$

By Markov's inequality, for all $x\geqslant 0$ and $\lambda\in(0,1/4)$,

$$\mathbb{P}\left(M_{T+1}^{m'}\geqslant V_{T+1}^{m'}\frac{\Psi(\lambda)}{\lambda}+\frac{x}{\lambda}\,\middle|\,\mathcal{H}_N\right)\leqslant e^{-x}. \tag{12}$$

Therefore, for all $x,u\geqslant 0$ and $\lambda\in(0,1/4)$,

$$\mathbb{P}\left(M_{T+1}^{m'}\geqslant \frac{\Psi(\lambda)}{\lambda}u+\frac{x}{\lambda}\quad\text{and}\quad V_{T+1}^{m'}\leqslant u\,\middle|\,\mathcal{H}_N\right)\leqslant e^{-x}.$$

To choose the optimal λ, we use Lemma 2 from [21].

Lemma 8. *[21, Lemma 2] Let a,b and x be positive constants and let us consider on $(0,1/b)$,*

$$g(\xi)=\frac{a\xi}{1-b\xi}+\frac{x}{\xi}.$$

Then $\min_{\xi\in(0,1/b)} g(\xi)=2\sqrt{ax}+bx$ and the minimum is achieved in $\xi(a,b,x)=\dfrac{\sqrt{x}}{\sqrt{x}b+\sqrt{a}}$.

For $a=\frac{u}{2}$ and $b=4$, Lemma 8 shows that for all $x,u\geqslant 0$,

$$\mathbb{P}\left(M_{T+1}^{m'}\geqslant \sqrt{2ux}+4x\quad\text{and}\quad V_{T+1}^{m'}\leqslant u\,\middle|\,\mathcal{H}_N\right)\leqslant e^{-x}.$$

Let us use a peeling argument similar to [21]:

Lemma 9. *Let X,V be real-valued random variables and α,b,v,w be positive numbers such that $V\in[w,v]$ a.s. and such that for all $x\geqslant 0$ and $u\in[w,v]$,*

$$\mathbb{P}(X\geqslant\sqrt{ux}+bx\quad\textit{and}\quad(1+\alpha)^{-1}u\leqslant V\leqslant u)\leqslant e^{-x},$$

then for any $x\geqslant 0$,

$$\mathbb{P}(X\geqslant\sqrt{(1+\alpha)Vx}+bx)\leqslant\left(1+\frac{\log(v/w)}{\log(1+\alpha)}\right)e^{-x}.$$

Proof. Let $v_0 = w$, $v_{d+1} = (1+\alpha)v_d$, and D the smallest integer such that $v_D \geqslant v$. For all $d \in [D]$ and $x \geqslant 0$,

$$\mathbb{P}\left(X \geqslant \sqrt{v_d x} + bx \quad \text{and} \quad v_{d-1} \leqslant V \leqslant v_d\right) \leqslant e^{-x}.$$

In particular, since $V \geqslant v_{d-1} = (1+\alpha)^{-1} v_d$ on this event,

$$\mathbb{P}\left(X \geqslant \sqrt{(1+\alpha)Vx} + bx \quad \text{and} \quad v_{d-1} \leqslant V \leqslant v_d\right) \leqslant e^{-x}.$$

Taking the union bound,

$$\mathbb{P}\left(X \geqslant \sqrt{(1+\alpha)Vx} + bx\right) \leqslant De^{-x},$$

and by definition $D \leqslant \dfrac{\log(v/w)}{\log(1+\alpha)} + 1$. □

We may apply Lemma 9 to $X = M^{m'}_{T+1}$ and $b = 4$. Since $V^{m'}_{T+1}$ does not have an obvious lower bound, we consider $V = 2(V^{m'}_{T+1} + \beta)$ for some $\beta > 0$ to be chosen later. We may therefore take $w = 2\beta$. For the upper bound v on V, by (11), since the Hellinger distance is upper bounded by 1, we may take $v = 2(\beta + 9(T - N + 1))$. With these choices, for any $\beta, \alpha, x > 0$,

$$\mathbb{P}\left(M^{m'}_{T+1} \geqslant \sqrt{2(1+\alpha)(V^{m'}_{T+1} + \beta)x} + 4x \,\middle|\, \mathcal{H}_N\right) \leqslant \left(\frac{\log\left(\frac{9(T-N+1)}{\beta} + 1\right)}{\log(1+\alpha)} + 1\right) e^{-x}.$$

For $\alpha = \sqrt{2}$,

$$\mathbb{P}\left(M^{m'}_{T+1} \geqslant \sqrt{5(V^{m'}_{T+1} + \beta)x} + 4x \,\middle|\, \mathcal{H}_N\right) \leqslant \left(2\log\left(\frac{9(T-N+1)}{\beta} + 1\right) + 1\right) e^{-x}. \quad (13)$$

By definition of $V^{m'}_{T+1}$ and the triangle inequality,

$$\begin{aligned} V^{m'}_{T+1} &= 9 \sum_{s=N}^{T} H\left(p^m_s(\cdot, X_s), p^{m'}_s(\cdot, X_s)\right)^2 \\ &\leqslant 18 \sum_{s=N}^{T} \left(H\left(p^\star_s(\cdot, X_s), p^m_s(\cdot, X_s)\right)^2 + H\left(p^\star_s(\cdot, X_s), p^{m'}_s(\cdot, X_s)\right)^2\right). \quad (14) \end{aligned}$$

We now use [29, Lemma 7.23] giving a connection between the Kullback-Leibler divergence $\mathrm{D_{KL}}$ and the Hellinger distance H.

Lemma 10. *[29, Lemma 7.23] Let P and Q be some probability measures. Then,*

$$\mathrm{D_{KL}}\left(P, \frac{P+Q}{2}\right) \geqslant (2\log 2 - 1)\, H^2(P, Q).$$

Moreover, whenever $P \ll Q$,

$$2H^2(P, Q) \leqslant \mathrm{D_{KL}}(P, Q).$$

Since $\dfrac{18}{2\log(2) - 1} \leqslant 48$,

$$\begin{aligned} V^{m'}_{T+1} &\leqslant 48 \sum_{s=N}^{T} \left(\mathrm{D_{KL}}\left(p^\star_s(\cdot, X_s), \frac{p^\star_s(\cdot, X_s) + p^{m'}_s(\cdot, X_s)}{2}\right) + \frac{1}{2}\mathrm{D_{KL}}\left(p^\star_s(\cdot, X_s), p^m_s(\cdot, X_s)\right)\right) \\ &=: 9W^{m'}_T. \quad (15) \end{aligned}$$

Let $\beta = 9(T - N + 1)y^2$, where $y > 0$ is to be chosen later. Replacing x by $x + \log(|\mathcal{M}|)$ leads to

$$\begin{aligned} &\mathbb{P}\left(\frac{M^{m'}_{T+1}}{T-N+1} \geqslant 3\sqrt{5\left(\frac{W^{m'}_T}{T-N+1} + y^2\right)\frac{x + \log(|\mathcal{M}|)}{T-N+1}} + 4\frac{x + \log(|\mathcal{M}|)}{T-N+1} \,\middle|\, \mathcal{H}_N\right) \\ &\qquad \leqslant \left(2\log\left(y^{-2} + 1\right) + 1\right) e^{-(x + \log(|\mathcal{M}|))}. \end{aligned}$$

Let $\kappa_1 \in (0, 1/(8\sqrt{5})]$, then, using $2\sqrt{ab} \leqslant \kappa_1 a + \kappa_1^{-1} b$ and taking $y^2 = \frac{x+\log(|\mathcal{M}|)}{(T-N+1)\log 2} \geqslant \frac{1}{T-N+1}$ since $x \geqslant 0$ and $|\mathcal{M}| \geqslant 2$,

$$\mathbb{P}\left(\frac{M_{T+1}^{m'}}{T-N+1} \geqslant \frac{3\sqrt{5}}{2}\kappa_1 \frac{W_T^{m'}}{T-N+1} + \left(4 + \frac{3\sqrt{5}}{2}\left(\frac{\kappa_1}{\log 2} + \kappa_1^{-1}\right)\right)\frac{x+\log(|\mathcal{M}|)}{T-N+1} \,\middle|\, \mathcal{H}_N\right) \leqslant (2\log(T-N+2)+1)\, e^{-(x+\log(|\mathcal{M}|))}. \quad (16)$$

By the union bound on all $m' \in \mathcal{M}$, the previous inequality holds with probability at least $1 - (2\log(T-N+2)+1)e^{-x}$ for all $m' \in \mathcal{M}$. It holds in particular for $\hat{m}$. Recall with (8) that,

$$\frac{1}{T-N+1}\sum_{s=N}^{T} \mathrm{D}_{\mathrm{KL}}\left(p_s^\star(\cdot, X_s), \frac{p_s^\star(\cdot, X_s) + p_s^{\hat{m}}(\cdot, X_s)}{2}\right) - U_m \leqslant \frac{1}{2(T-N+1)}\sum_{s=N}^{T} \mathrm{D}_{\mathrm{KL}}\left(p_s^\star(\cdot, X_s), p_s^m(\cdot, X_s)\right) + \frac{M_{T+1}^{\hat{m}}}{T-N+1}. \quad (17)$$

Plugging (15) and (16) in (17) leads to, conditionally on $\mathcal{H}_N$, with probability at least $1-(2\log(T-N+2)+1)e^{-x}$

$$\frac{(1-C_{\kappa_1})}{T-N+1}\sum_{s=N}^{T} \mathrm{D}_{\mathrm{KL}}\left(p_s^\star(\cdot, X_s), \frac{p_s^\star(\cdot, X_s) + p_s^{\hat{m}}(\cdot, X_s)}{2}\right) - U_m \leqslant \frac{(1+C_{\kappa_1})}{T-N+1}\sum_{s=N}^{T} \frac{1}{2}\mathrm{D}_{\mathrm{KL}}\left(p_s^\star(\cdot, X_s), p_s^m(\cdot, X_s)\right) + C'_{\kappa_1}\frac{x+\log(|\mathcal{M}|)}{T-N+1},$$

where

- $C_{\kappa_1} = 8\sqrt{5}\kappa_1$,
- $C'_{\kappa_1} = 4 + \frac{3\sqrt{5}}{2}\left(\frac{\kappa_1}{\log 2} + \kappa_1^{-1}\right) \leqslant 13\kappa_1^{-1} = \frac{104\sqrt{5}}{C_{\kappa_1}}$,

Integrating on $x \geqslant 0$ and noting that $\mathbb{E}[U_m|\mathcal{H}_N] = 0$ leads to, for all $m \in \mathcal{M}$,

$$\frac{(1-C_{\kappa_1})}{T-N+1}\mathbb{E}\left[\sum_{s=N}^{T} \mathrm{D}_{\mathrm{KL}}\left(p_s^\star(\cdot, X_s), \frac{p_s^\star(\cdot, X_s) + p_s^{\hat{m}}(\cdot, X_s)}{2}\right)\,\middle|\, \mathcal{H}_N\right] \leqslant \frac{(1+C_{\kappa_1})}{T-N+1}\mathbb{E}\left[\frac{1}{2}\sum_{s=N}^{T} \mathrm{D}_{\mathrm{KL}}\left(p_s^\star(\cdot, X_s), p_s^m(\cdot, X_s)\right)\,\middle|\, \mathcal{H}_N\right] + \frac{104\sqrt{5}}{C_{\kappa_1}}\frac{2\log(T-N+2)+1+\log(|\mathcal{M}|)}{T-N+1}.$$

To conclude $\kappa = \frac{\kappa_1}{8\sqrt{5}}$, so that $C_{\kappa_1} = \kappa$. $\square$

C.2 Proof of Section 3

Proof of Proposition 2. The proof is straightforward with the definition of $p_{\theta^m,t}^m$ in (1). Let $\theta^m, \delta^m \in \Theta^m$, $t \leqslant T_\varepsilon$, $x \in \mathcal{X}$, and $k \in [K]$.

$$p_{\theta^m,t}^m(k,x) = \sum_{C \in \mathcal{P}_m} \pi_{C,T_t^C}^{\theta_C}(k)\mathbf{1}_{x\in C} \geqslant \sum_{C \in \mathcal{P}_m} \varepsilon \mathbf{1}_{x \in C} = \varepsilon.$$

For the second part of the proof, it holds that, almost surely, for all $t \leqslant T_\varepsilon$

$$\left|\log\left(\frac{p_{\delta^m,t}^m(k,x)}{p_{\theta^m,t}^m(k,x)}\right)\right| = \sum_{C\in\mathcal{P}_m}\left|\log\left(\frac{\pi_{C,T_t^C}^{\delta_C}(k)}{\pi_{C,T_t^C}^{\theta_C}(k)}\right)\right|\mathbf{1}_{x\in C} \leqslant L_\varepsilon \sum_{C\in\mathcal{P}_m}\|\delta_C - \theta_C\|_2\,\mathbf{1}_{x\in C} \leqslant L_\varepsilon \sup_{C\in\mathcal{P}_m}\|\delta_C - \theta_C\|_2.$$

$\square$

Proof of Theorem 3. Our goal is to apply [6].

Assumption 1 of [6] is satisfied for $n = T_\varepsilon$ since with Proposition 2, there exists $\varepsilon > 0$ such that a.s., for all $t \in [T_\varepsilon]$, for all $k \in [K]$, $p_t^\star(k, X_t) \in [\varepsilon, 1]$ and for all $m \in \mathcal{M}$ and all $\theta^m \in \Theta^{D_m}$, $p^m_{\theta^m,t}(k, X_t) \in [\varepsilon, 1]$.

Assumption 2 of [6] is satisfied since with Proposition 2, there exists a positive constants L_ε such that a.s., for all $t \in [T_\varepsilon]$, for all $m \in \mathcal{M}$ and all $\delta^m, \theta^m \in \Theta^{D_m}$,

$$\sup_{k\in[K]} \left| \log \left(\frac{p^m_{\delta^m,t}(k, X_t)}{p^m_{\theta^m,t}(k, X_t)} \right) \right| \leqslant L_\varepsilon \sup_{C\in\mathcal{P}_m} \|\delta_C - \theta_C\|_2$$

and by Assumption, for all $\theta^m, \delta^m \in \Theta^{D_m}$

$$\sup_{C\in\mathcal{P}_m} \|\delta_C - \theta_C\|_2 \leqslant \sqrt{d}(R - r).$$

Note in particular that the Lipschitz constant in Proposition 2 does not depend on m.

Assumption 3 in [6] is always satisfied because the set of actions $[K]$ is finite.

Setting $A_\varepsilon = L_\varepsilon \sqrt{d}(R - r) + 2 \log(\varepsilon^{-1})$, Corollary 2 in [6] simply reads as follows. There exist positive numerical constants C and C' such that the following holds. Assume that

$$\Sigma_\varepsilon = \log(A_\varepsilon) \sum_{m\in\mathcal{M}} e^{-D_m} < +\infty.$$

Let $\kappa \in (0, 1]$. If for all $m \in \mathcal{M}$,

$$\text{pen}(m) \geqslant \frac{C}{\kappa} A_\varepsilon^2 \log(\varepsilon^{-1})^{3/2} \log(T_\varepsilon A_\varepsilon)^2 \frac{D_m}{T_\varepsilon},$$

then,

$$\begin{aligned}
&\frac{1-\kappa}{T_\varepsilon} \sum_{t=1}^{T_\varepsilon} \mathbb{E}\left[\mathrm{D_{KL}} \left(p_t^\star(\cdot, X_t), p^{\hat{m}}_{\hat{\theta}^{\hat{m}},t}(\cdot, X_t) \right) \right] \\
&\quad \leqslant \inf_{m\in\mathcal{M}} \left((1+\kappa) \inf_{\theta\in\Theta^m} \frac{1}{T_\varepsilon} \sum_{t=1}^{T_\varepsilon} \mathbb{E}\left[\mathrm{D_{KL}} \left(p_t^\star(\cdot, X_t), p^m_{\theta^m,t}(\cdot, X_t) \right) \right] + 2\,\text{pen}(m) \right) \\
&\qquad + \frac{18C'}{\kappa} A_\varepsilon \Sigma_\varepsilon \log(\varepsilon^{-1})^{3/2} \log(T_\varepsilon A_\varepsilon)^2 \frac{\log(T_\varepsilon)}{T_\varepsilon}.
\end{aligned}$$

□

C.3 Proof of Section 4.1

Proof of Proposition 4. Let $\theta_C \in \Theta_C$. Write $\theta_{C,T} = (\eta_{C,T}, \gamma_{C,T}) = \theta_C/\sqrt{T} \in \Theta$. To ease the notations in the proof, we remove the C from the notations. θ_C becomes θ and $\theta_{C,T}$ becomes θ_T. In the same way, $\theta_T = (\eta_T, \gamma_T)$ now.

Let $t \in F_T(C)$. Then,

$$\begin{aligned}
\pi^\theta_{C,T_t^C+1}(k) = &\frac{\pi^\theta_{C,T_t^C}(k) e^{-\eta_T g_{k,t}/(\gamma_T + \pi^\theta_{C,T_t^C}(k))} \mathbf{1}_{A_t=k}}{(1 - \pi^\theta_{C,T_t^C}(k)) + \pi^\theta_{C,T_t^C}(k) e^{-\eta_T g_{k,t}/(\gamma_T + \pi^\theta_{C,T_t^C}(k))}} \\
&+ \sum_{\substack{j=1\\ j\neq k}}^{K} \frac{\pi^\theta_{C,T_t^C}(k) \mathbf{1}_{A_t=j}}{(1 - \pi^\theta_{C,T_t^C}(j)) + \pi^\theta_{C,T_t^C}(j) e^{-\eta_T g_{j,t}/(\gamma_T + \pi^\theta_{C,T_t^C}(j))}}.
\end{aligned}$$

For any $q \in [0, 1]$, since $g_{k,t} \in [0, 1]$, $1 - q + q e^{-\eta g_{k,t}/(q+\gamma)} \leqslant 1$. Therefore,

$$\pi^\theta_{C,T_t^C+1}(k) \geqslant \pi^\theta_{C,T_t^C}(k) e^{-\eta_T g_{k,t}/(\gamma_T + \pi^\theta_{C,T_t^C}(k))} \mathbf{1}_{A_t=k} + \sum_{\substack{j=1\\ j\neq k}}^{K} \pi^\theta_{C,T_t^C}(k) \mathbf{1}_{A_t=j}.$$

Since $e^{-\eta_T g_{k,t}/(\gamma_T+\pi^\theta_{C,T^C_t}(k))} \leqslant 1$,

$$\begin{aligned}
&\pi^\theta_{C,T^C_t+1}(k)\\
&\geqslant \pi^\theta_{C,T^C_t}(k) e^{-\eta_T g_{k,t}/(\gamma_T+\pi^\theta_{C,T^C_t}(k))}\mathbf{1}_{A_t=k} + \sum_{\substack{j=1\\ j\neq k}}^K \pi^\theta_{C,T^C_t}(k) e^{-\eta_T g_{k,t}/(\gamma_T+\pi^\theta_{C,T^C_t}(k))}\mathbf{1}_{A_t=j}\\
&= \pi^\theta_{C,T^C_t}(k) e^{-\eta_T g_{k,t}/(\gamma_T+\pi^\theta_{C,T^C_t}(k))} \geqslant \pi^\theta_{C,T^C_t}(k) e^{-\eta_T/(\gamma_T+\pi^\theta_{C,T^C_t}(k))}
\end{aligned}$$

since $g_{k,t} \in [0,1]$. Then,

$$\pi^\theta_{C,T^C_t+1}(k) \geqslant \pi^\theta_{C,T^C_t}(k)\left(1 - \frac{\eta_T g_{k,t}}{\gamma_T + \pi^\theta_{C,T^C_t}(k)}\right) \geqslant \pi^\theta_{C,T^C_t}(k) - \eta_T.$$

Summing for all $s \in F_t(C)$, since $\pi^\theta_{k,1} = \frac{1}{K}$,

$$\pi^\theta_{C,T^C_t}(k) \geqslant \frac{1}{K} - \eta_T T^C_t.$$

Note that $T^C_t \leqslant t \leqslant T_\varepsilon$. Since, $T_\varepsilon \leqslant \left\lfloor \left(\frac{1}{K} - \varepsilon\right)\frac{\sqrt{T}}{R}\right\rfloor$, for all $t \leqslant T_\varepsilon$ and $1 \leqslant k \leqslant K$,

$$\varepsilon \leqslant \frac{1}{K} - \frac{R}{\sqrt{T}} T_\varepsilon \leqslant \frac{1}{K} - \eta_T T_\varepsilon \leqslant \frac{1}{K} - \eta_T t \leqslant \pi^\theta_{C,T^C_t}(k).$$

For the second part of the proof, let $\theta = (\eta, \gamma), \theta' = (\eta', \gamma') \in \Theta_C$. For $t \geqslant 2$ Let $h^\theta_{j,t} = \eta_T \sum_{s\in F_t(C)} \hat{g}^\theta_{j,s}$. Then $\pi^\theta_{C,T^C_t} = \mathrm{softmax}(h^\theta_{\cdot,t})$. The function softmax is 1-Lipschitz with respect to the $\|\cdot\|_2$-norm in $\mathbb{R}^K$ (see [19] for a proof). Therefore,

$$\|\pi^\theta_{C,T^C_t} - \pi^{\theta'}_{C,T^C_t}\|_2 \leqslant \|h^\theta_{\cdot,t} - h^{\theta'}_{\cdot,t}\|_2.$$

Since $g_{j,s} \in [0,1]$, by the triangle inequality

$$\|\pi^\theta_{C,T^C_t} - \pi^{\theta'}_{C,T^C_t}\|_2 \leqslant \sum_{s\in F_t(C)} \left\| \left(\frac{\eta_T}{\gamma_T + \pi^\theta_{C,T^C_s}(.)} - \frac{\eta'_T}{\gamma'_T + \pi^{\theta'}_{C,T^C_s}(.)}\right)\mathbf{1}_{A_s=\cdot}\right\|_2.$$

Again, using the triangle inequality,

$$\begin{aligned}
\|\pi^\theta_{C,T^C_t} - \pi^{\theta'}_{C,T^C_t}\|_2 \leqslant& \sum_{s\in F_t(C)} \left\| \left(\frac{\eta_T}{\gamma_T + \pi^\theta_{C,T^C_s}} - \frac{\eta'_T}{\gamma'_T + \pi^\theta_{C,T^C_s}}\right)\mathbf{1}_{A_s=\cdot}\right\|_2\\
&+ \sum_{s\in F_t(C)} \left\| \left(\frac{\eta'_T}{\gamma'_T + \pi^\theta_{C,T^C_s}} - \frac{\eta'_T}{\gamma'_T + \pi^{\theta'}_{C,T^C_s}}\right)\mathbf{1}_{A_s=\cdot}\right\|_2. \qquad (18)
\end{aligned}$$

For $1 \geqslant q \geqslant \varepsilon$, let $f : (x_1, x_2) \in [0, R_T] \times [0, R_T] \mapsto \dfrac{x_1}{x_2 + q}$ where $R_T = \frac{R}{\sqrt{T}}$. The function f is continuously differentiable, and

$$\nabla f = \frac{1}{(x_2+q)^2}\begin{pmatrix} x_2 + q \\ -x_1 \end{pmatrix}.$$

The ℓ^2-norm of the gradient can be upper bounded by

$$\|\nabla f\|_2 \leqslant \frac{1}{\varepsilon^2}\sqrt{R_T^2 + \varepsilon^2} =: c_\varepsilon$$

By the mean value theorem, for all $k \in [K]$

$$\left| \frac{\eta_T}{\gamma_T + \pi^\theta_{C,T^C_s}(k)} - \frac{\eta'_T}{\gamma'_T + \pi^\theta_{C,T^C_s}(k)} \right| \leqslant c_\varepsilon \|\theta_T - \theta'_T\|_2.$$

As a result,

$$\left\|\left(\frac{\eta_T}{\gamma_T+\pi^\theta_{C,T^C_s}}-\frac{\eta'_T}{\gamma'_T+\pi^\theta_{C,T^C_s}}\right)\mathbf{1}_{A_s=\cdot}\right\|_2^2=\sum_{k=1}^K\left(\frac{\eta_T}{\gamma_T+\pi^\theta_{C,T^C_s}(k)}-\frac{\eta'_T}{\gamma'_T+\pi^\theta_{C,T^C_s}(k)}\right)^2\mathbf{1}_{A_s=k}$$
$$\leqslant c_\varepsilon^2\|\theta_T-\theta'_T\|_2^2\sum_{k=1}^K\mathbf{1}_{A_s=k}$$
$$=c_\varepsilon^2\|\theta_T-\theta'_T\|_2^2$$

Therefore,

$$\sum_{s\in F_t(C)}\left\|\left(\frac{\eta_T}{\gamma_T+\pi^\theta_{C,T^C_s}}-\frac{\eta'_T}{\gamma'_T+\pi^\theta_{C,T^C_s}}\right)\mathbf{1}_{A_s=\cdot}\right\|_2$$
$$\leqslant\sum_{s\in F_t(C)}c_\varepsilon\|\theta_T-\theta'_T\|_2=T^C_t c_\varepsilon\|\theta_T-\theta'_T\|_2\leqslant T_\varepsilon c_\varepsilon\|\theta_T-\theta'_T\|_2. \tag{19}$$

For $(\eta,\gamma)\in[0,R_T]\times[0,R_T]$, let $g:q\in[\varepsilon,1]\mapsto\dfrac{\eta}{\gamma+q}$. The function g is continuously differentiable, and

$$0\leqslant f'(q)=\frac{\eta}{(\gamma+q)^2}\leqslant\frac{R_T}{\varepsilon^2}.$$

By the mean value theorem, for all $k\in[K]$,

$$\left|\frac{\eta'_T}{\gamma'_T+\pi^\theta_{C,T^C_s}(k)}-\frac{\eta'_T}{\gamma'_T+\pi^{\theta'}_{C,T^C_s}}(k)\right|\leqslant\frac{R_T}{\varepsilon^2}\left|\pi^\theta_{C,T^C_s}(k)-\pi^{\theta'}_{C,T^C_s}(k)\right|.$$

Therefore,

$$\left\|\left(\frac{\eta'_T}{\gamma'_T+\pi^\theta_{C,T^C_s}}-\frac{\eta'_T}{\gamma'_T+\pi^{\theta'}_{C,T^C_s}}\right)\mathbf{1}_{A_s=\cdot}\right\|_2^2=\sum_{k=1}^K\left(\frac{\eta'_T}{\gamma'_T+\pi^\theta_{C,T^C_s}(k)}-\frac{\eta'_T}{\gamma'_T+\pi^{\theta'}_{C,T^C_s}}\right)^2\mathbf{1}_{A_s=k}$$
$$\leqslant\frac{R_T^2}{\varepsilon^4}\sum_{k=1}^K\left|\pi^\theta_{C,T^C_s}(k)-\pi^{\theta'}_{C,T^C_s}(k)\right|^2\mathbf{1}_{A_s=k}$$
$$\leqslant\frac{R_T^2}{\varepsilon^4}\left(\left\|\pi^\theta_{C,T^C_s}-\pi^{\theta'}_{C,T^C_s}\right\|_2\right)^2.$$

Thus,

$$\sum_{s\in F_t(C)}\left\|\left(\frac{\eta'_T}{\gamma'_T+\pi^\theta_{C,T^C_s}}-\frac{\eta'_T}{\gamma'_T+\pi^{\theta'}_{C,T^C_s}}\right)\mathbf{1}_{A_s=\cdot}\right\|_2\leqslant\frac{R_T}{\varepsilon^2}\sum_{s\in F_t(C)}\left\|\pi^\theta_{C,T^C_s}-\pi^{\theta'}_{C,T^C_s}\right\|_2 \tag{20}$$

Plugging Equations (19) and (20) in Equation (18)

$$\|\pi^\theta_{C,T^C_t}-\pi^{\theta'}_{C,T^C_t}\|_2\leqslant T_\varepsilon c_\varepsilon\|\theta_T-\theta'_T\|_2+\frac{R_T}{\varepsilon^2}\sum_{s\in F_t(C)}\left\|\pi^\theta_{C,T^C_s}-\pi^{\theta'}_{C,T^C_s}\right\|_2,$$

Using the discrete Gronwall Lemma [14] leads to, for all $t\leqslant T_\varepsilon$,

$$\|\pi^\theta_{C,T^C_t}-\pi^{\theta'}_{C,T^C_t}\|_2\leqslant T_\varepsilon c_\varepsilon\|\theta_T-\theta'_T\|_2\prod_{s\in F_t(C)}\left(1+\frac{R_T}{\varepsilon^2}\right)$$
$$\leqslant T_\varepsilon c_\varepsilon\|\theta_T-\theta'_T\|_2\exp\left(\frac{R_TT_\varepsilon}{\varepsilon^2}\right).$$

But, since $\frac{1}{K}-\varepsilon\leqslant 1$,

$$T_\varepsilon\leqslant\left(\frac{1}{K}-\varepsilon\right)\frac{\sqrt{T}}{R}\leqslant\frac{\sqrt{T}}{R}.$$

Therefore, $R_T T_\varepsilon \leqslant 1$ and for $t \leqslant T_\varepsilon$,

$$\|\pi^{\theta}_{C,T^C_t} - \pi^{\theta'}_{C,T^C_t}\|_2 \leqslant \frac{c_\varepsilon}{R} e^{1/\varepsilon^2} \|\theta - \theta'\|_2.$$

To conclude note that $\log$ is $1/\varepsilon$-Lipschitz on $[\varepsilon, 1]$ and that

$$\sup_{k \in [K]} \left| \log \left(\frac{\pi^{\delta_C}_{C,T^C_t}(k)}{\pi^{\theta_C}_{C,T^C_t}(k)} \right) \right| \leqslant \frac{1}{\varepsilon} \|\pi^{\theta}_{C,T^C_t} - \pi^{\theta'}_{C,T^C_t}\|_2.$$

□

C.4 Proof of Section 4.2

Proof of Proposition 5. We take the same notations as the previous section. The updated probability can be written

$$\pi^{\theta}_{C,T^C_t+1}(k) = \frac{\pi^{\theta}_{C,T^C_t}(k) e^{\theta_T \left(\mathbf{1}_{A_t = k} - \pi^{\theta}_{C,T^C_t}(k)\right) g_{A_t,t}}}{\pi^{\theta}_{C,T^C_t}(A_t) e^{\theta_T (1 - \pi^{\theta}_{C,T^C_t}(A_t)) g_{A_t,t}} + \sum_{j \neq A_t} \pi^{\theta}_{C,T^C_t}(j) e^{-\theta_T \pi^{\theta}_{C,T^C_t}(j) g_{A_t,t}}}.$$

Therefore,

$$\begin{aligned}
\pi^{\theta}_{C,T^C_t+1}(k) &\geqslant \frac{\pi^{\theta}_{C,T^C_t}(k) e^{\theta_T \left(\mathbf{1}_{A_t = k} - \pi^{\theta}_{C,T^C_t}(k)\right) g_{A_t,t}}}{\pi^{\theta}_{C,T^C_t}(A_t) e^{\theta_T (1 - \pi^{\theta}_{C,T^C_t}(A_t)) g_{A_t,t}} + \sum_{j \neq A_t} \pi^{\theta}_{C,T^C_t}(j)} \\
&\geqslant \frac{\pi^{\theta}_{C,T^C_t}(k) e^{-\theta_T}}{\pi^{\theta}_{C,T^C_t}(A_t) e^{\theta_T} + 1 - \pi^{\theta}_{C,T^C_t}(A_t)} \\
&\geqslant \pi^{\theta}_{C,T^C_t}(k) e^{-2\theta_T}
\end{aligned}$$

where

- the first inequality holds because $e^{-\theta_T \pi^{\theta}_{C,T^C_t}(j) g_{A_t,t}} \leqslant 1$,
- the second inequality holds because $g_{j,t} \in (0,1)$, for $j \in [K]$,
- the last inequality holds because $\pi^{\theta}_{C,T^C_t}(A_t) e^{\theta_T} + 1 - \pi^{\theta}_{C,T^C_t}(A_t) \leqslant e^{\theta_T}$.

Thus, for all $t \leqslant T_\varepsilon$,

$$\pi^{\theta}_{C,T^C_t}(k) \geqslant \frac{1}{K} e^{-2\theta_T T^C_t}.$$

Since $T^C_t \leqslant t$ and since by definition, $T_\varepsilon \leqslant \log\left(\sqrt{\frac{1}{K\varepsilon}}\right) \frac{\sqrt{T}}{R}$, it holds that for all $t \leqslant T_\varepsilon$,

$$\pi^{\theta}_{C,T^C_t}(k) \geqslant \frac{1}{K} e^{-2\theta_T t} \geqslant \frac{1}{K} e^{-2R_T t} \geqslant \frac{1}{K} e^{-2\frac{R}{\sqrt{T}} \log\left(\sqrt{\frac{1}{K\varepsilon}}\right) \frac{\sqrt{T}}{R}} \geqslant \frac{1}{K} e^{-\log\left(\frac{1}{K\varepsilon}\right)} \geqslant \varepsilon.$$

For the second part of the proof, for $t \geqslant 2$ and $j \in [K]$, let $h^{\theta}_{j,T^C_t} = \theta_T \sum_{s \in F_t(C)} \hat{g}^{\theta}_{j,s}$. Then $\pi^{\theta}_{C,T^C_t} = \mathrm{softmax}(h^{\theta}_{\cdot,t})$. The function softmax is 1-Lipschitz with respect to the $\|\cdot\|_2$-norm in $\mathbb{R}^K$ (see [19] for a proof). Therefore,

$$\|\pi^{\delta}_{C,T^C_t} - \pi^{\theta}_{C,T^C_t}\|_2 \leqslant \|h^{\delta}_{\cdot,t} - h^{\theta}_{\cdot,t}\|_2.$$

Then,

$$\begin{aligned}
\|h^{\delta}_{\cdot,t} - h^{\theta}_{\cdot,t}\|_2 &\leqslant |\delta_T - \theta_T| \sum_{s \in F_t(C)} \|\hat{g}^{\delta}_{j,s}\|_2 + \theta_T \sum_{s \in F_t(C)} g_{A_s,s} \|\pi^{\delta}_{C,T^C_s} - \pi^{\theta}_{C,T^C_s}\|_2 \\
&\leqslant \sqrt{2} T^C_t |\delta_T - \theta_T| + \theta_T \sum_{s \in F_t(C)} \|\pi^{\delta}_{C,T^C_s} - \pi^{\theta}_{C,T^C_s}\|_2 \\
&\leqslant \sqrt{2} T_\varepsilon |\delta_T - \theta_T| + \theta_T \sum_{s \in F_t(C)} \|\pi^{\delta}_{C,T^C_s} - \pi^{\theta}_{C,T^C_s}\|_2.
\end{aligned}$$

where

- the first inequality holds because of the triangle inequality,
- the second inequality holds because for all $j \in [K]$, $g_{j,s} \in [0,1]$ and

$$\|\hat{g}^{\delta}_{j,s}\|_2^2 = (1 - \pi^{\delta}_{C,T^C_s}(A_s))^2 + \sum_{j \neq A_s} (\pi^{\delta}_{C,T^C_s}(j))^2 \leqslant 2,$$

- the last inequality holds because $T^C_t \leqslant T_\varepsilon$.

By the discrete Gronwall Lemma [14], for all $t \leqslant T_\varepsilon$

$$\|\pi^{\delta}_{C,T^C_t} - \pi^{\theta}_{C,T^C_t}\|_2 \leqslant \sqrt{2}|\delta_T - \theta_T|T_\varepsilon \prod_{s \in F_t(C)} (1+\theta_T) \leqslant \sqrt{2}|\delta_T - \theta_T|T_\varepsilon e^{\theta_T T_\varepsilon}.$$

Since $T_\varepsilon \leqslant \left\lfloor \log\left(\sqrt{\frac{1}{K\varepsilon}}\right) \frac{\sqrt{T}}{R} \right\rfloor$, $\theta_T T_\varepsilon \leqslant R_T T_\varepsilon \leqslant \log\left(\sqrt{\frac{1}{K\varepsilon}}\right)$, therefore,

$$\|\pi^{\delta}_{C,T^C_t} - \pi^{\theta}_{C,T^C_t}\|_2 \leqslant \frac{\sqrt{2}}{R} \frac{\log\left(\sqrt{\frac{1}{K\varepsilon}}\right)}{\sqrt{K\varepsilon}} |\delta - \theta|.$$

Finally, log is $\frac{1}{\varepsilon}$-Lipschitz on $[\varepsilon, 1]$. Thus, for all $k \in [K]$, $t \leqslant T_\varepsilon$,

$$\left|\log\left(\frac{\pi^{\delta}_{C,T^C_t}(k)}{\pi^{\theta}_{C,T^C_t}(k)}\right)\right| \leqslant \frac{1}{\varepsilon}|\pi^{\delta}_{C,T^C_t}(k) - \pi^{\theta}_{C,T^C_t}(k)| \leqslant \frac{\sqrt{2}}{R\varepsilon} \frac{\log\left(\sqrt{\frac{1}{K\varepsilon}}\right)}{\sqrt{K\varepsilon}} |\delta - \theta|.$$

□

C.5 Proof of Section B

Let us recall that $|F| = \max_{m \in \mathcal{M}} |E_m|$. For this Section, we drop the dependence m of the model and simply write E and θ generic set of policies and parameter in $[r, R]$.

Proof of Proposition 6. For any $\theta \in [r, R]$, write $\theta_T = \theta/\sqrt{T}$. Assume that Assumption 3 holds. Let's write $q^{\theta}_{E,t+1}$ as

$$q^{\theta}_{E,t+1}(j) = \frac{q^{\theta}_{E,t}(j) e^{-\theta_T \hat{y}^{\theta}_{j,t}}}{\sum_{i \in E} q^{\theta}_{E,t}(i) e^{-\theta_T \hat{y}^{\theta}_{i,t}}}$$

Since $q^{\theta}_{E,t}$ is a probability distribution over the experts,

$$\sum_{i \in E} q^{\theta}_{E,t}(i) e^{-\theta_T \hat{y}^{\theta}_{i,t}} \leqslant \sum_{i \in E} q^{\theta}_{E,t}(i) = 1.$$

Therefore, $q^{\theta}_{E,t+1}(j) \geqslant q^{\theta}_{E,t}(j) e^{-\theta_T \hat{y}^{\theta}_{j,t}}$. By definition,

$$\hat{y}^{\theta}_{j,t} = \sum_{k=1}^{K} \xi_{j,t}(k) \frac{g_{k,t}}{\pi^{\theta}_{E,t}(k)} \mathbf{1}_{A_t = k} = \xi_{j,t}(A_t) \frac{g_{A_t,t}}{\pi^{\theta}_{E,t}(A_t)}.$$

Using that $e^{-x} \geqslant 1 - x$ for any $x \geqslant 0$, leads to

$$q^{\theta}_{E,t+1}(j) \geqslant q^{\theta}_{E,t}(j) \left(1 - \theta_T \xi_{j,t}(A_t) \frac{g_{A_t,t}}{\pi^{\theta}_{E,t}(A_t)}\right).$$

Since $g_{A_t,t} \in [0,1]$ and $q^{\theta}_{E,t}(j)\xi_{j,t}(A_t) \leqslant \pi^{\theta}_{E,t}(A_t)$,

$$q^{\theta}_{E,t+1}(j) \geqslant q^{\theta}_{E,t}(j) - \theta_T.$$

Summing for all s from 1 to t,

$$q^{\theta}_{E,t}(j) \geqslant \frac{1}{|E|} - \theta_T t.$$

1016 Since

$$T_\varepsilon = \left\lfloor \left(\frac{1}{|F|} - \frac{\varepsilon}{\rho} \right) \frac{\sqrt{T}}{R} \right\rfloor \wedge T \qquad \text{and} \qquad |F| \geqslant |E|,$$

1017 it holds that,

$$T_\varepsilon \leqslant \left\lfloor \left(\frac{1}{|E|} - \frac{\varepsilon}{\rho} \right) \frac{\sqrt{T}}{R} \right\rfloor \wedge T.$$

1018 Therefore, for all $t \leqslant T_\varepsilon$,

$$q^\theta_{E,t}(j) \geqslant \frac{1}{|E|} - \frac{R}{\sqrt{T}} t \geqslant \frac{1}{|E|} - \frac{R}{\sqrt{T}} \left(\frac{1}{|E|} - \frac{\varepsilon}{\rho} \right) \frac{\sqrt{T}}{R} = \frac{\varepsilon}{\rho}.$$

1019 Finally, for all $t \leqslant T_\varepsilon$, for all $k \in [K]$,

$$\pi^\theta_{E,t}(k) = \sum_{j \in E} q^\theta_{E,t}(j) \xi_{j,t}(k) \geqslant \frac{\varepsilon}{\rho} \sum_{j \in E} \xi_{j,t}(k) = \varepsilon.$$

1020 For the second part of the proof, let $\eta, \delta \in [r, R]$, and write $\eta_T = \eta/\sqrt{T}$ and likewise for δ_T.
1021 For $t \geqslant 2$, let $g^\eta_{j,t} = \eta_T \sum_{s=1}^{t-1} \hat{y}^\eta_{j,s}$. Then, $q^\eta_{E,t} = \mathrm{softmax}(g^\eta_t)$ where $g^\eta_t = (g^\eta_{j,t})_{j \in E}$. Since the
1022 function softmax is 1-Lipschitz with respect to the $\|\cdot\|_2$-norm in $\mathbb{R}^{|E|}$,

$$\|q^\eta_{E,t} - q^\delta_{E,t}\|_2 \leqslant \|g^\eta_t - g^\delta_t\|_2.$$

1023 Therefore, by the triangle inequality,

$$\begin{aligned} \|q^\eta_{E,t} - q^\delta_{E,t}\|_2 &\leqslant \sum_{s=1}^{t-1} \|\eta_T \hat{y}^\eta_{j,s} - \delta_T \hat{y}^\delta_{j,s}\|_2 \\ &\leqslant \sum_{s=1}^{t-1} \left(|\eta_T - \delta_T| \|\hat{y}^\eta_{j,s}\|_2 + \delta_T \|\hat{y}^\eta_{j,s} - \hat{y}^\delta_{j,s}\|_2 \right). \end{aligned} \tag{21}$$

1024 Since $\xi_{j,t}$ is a probability distribution,

$$\begin{aligned} \|\hat{y}^\eta_{j,t}\|_2^2 = \sum_{j \in E} (\hat{y}^\eta_{j,t})^2 = \sum_{j \in E} \left(\sum_{k=1}^K \xi_{j,t}(k) \frac{g_{k,t}}{\pi^\eta_{E,t}(k)} \mathbf{1}_{A_t = k} \right)^2 \\ = \sum_{j \in E} \left(\xi_{j,t}(A_t) \frac{g_{A_t,t}}{\pi^\eta_{E,t}(A_t)} \right)^2 \leqslant |E| \left(\frac{g_{A_t,t}}{\pi^\eta_{E,t}(A_t)} \right)^2. \end{aligned}$$

1025 Since $g^\eta_{A_t,t} \in [0,1]$ and $\pi^\eta_{E,t}(A_t) \geqslant \varepsilon$ for all $t \leqslant T_\varepsilon$,

$$\|\hat{y}^\eta_{j,t}\|_2 \leqslant \frac{\sqrt{|E|}}{\varepsilon}. \tag{22}$$

1026 Similarly,

$$\begin{aligned} \|\hat{y}^\eta_{j,s} - \hat{y}^\delta_{j,s}\|_2^2 &= \sum_{j \in E} \left(\sum_{k=1}^K \xi_{j,t}(k) g_{k,t} \left(\frac{1}{\pi^\eta_{E,t}(k)} - \frac{1}{\pi^\delta_{E,t}(k)} \right) \mathbf{1}_{A_t = k} \right)^2 \\ &\leqslant \sum_{j \in E} \left(\sum_{k=1}^K \xi_{j,t}(k) \left(\frac{1}{\pi^\eta_{E,t}(k)} - \frac{1}{\pi^\delta_{E,t}(k)} \right) \mathbf{1}_{A_t = k} \right)^2. \end{aligned}$$

1027 Thus, for all $t \leqslant T_\varepsilon$,

$$\|\hat{y}^\eta_{j,t} - \hat{y}^\delta_{j,t}\|_2^2 \leqslant \frac{1}{\varepsilon^4} \sum_{j \in E} \left(\sum_{k=1}^K \xi_{j,t}(k) \left(\pi^\eta_{E,t}(k) - \pi^\delta_{E,t}(k) \right) \mathbf{1}_{A_t = k} \right)^2.$$

Since $\xi_{j,t}(k) \leqslant 1$,

$$\|\hat{y}^{\eta}_{j,t} - \hat{y}^{\delta}_{j,t}\|_2 \leqslant \frac{\sqrt{|E|}}{\varepsilon^2} \|\pi^{\eta}_{E,t} - \pi^{\delta}_{E,t}\|_2. \tag{23}$$

Injecting (22) and (23) in (21) leads to

$$\|q^{\eta}_{E,t} - q^{\delta}_{E,t}\|_2 \leqslant \frac{\sqrt{|E|}}{\varepsilon} |\eta_T - \delta_T|(t-1) + \delta_T \frac{\sqrt{|E|}}{\varepsilon^2} \sum_{s=1}^{t-1} \|\pi^{\eta}_{E,s} - \pi^{\delta}_{E,s}\|_2.$$

On the other hand,

$$\begin{aligned}
\|\pi^{\eta}_{E,t} - \pi^{\delta}_{E,t}\|_2^2 = \sum_{k=1}^{K} (\pi^{\eta}_{k,t} - \pi^{\delta}_{k,t})^2 &= \sum_{k=1}^{K} \left(\sum_{j \in E} \xi_{j,t}(k) (q^{\eta}_{E,t}(j) - q^{\delta}_{E,t}(j)) \right)^2 \\
&\leqslant \sum_{k=1}^{K} \sum_{j \in E} \xi_{j,t}(k)^2 \sum_{j \in E} (q^{\eta}_{E,t}(j) - q^{\delta}_{E,t}(j))^2 \\
&\leqslant |E| \|q^{\eta}_{E,t} - q^{\delta}_{E,t}\|_2^2 \\
&\leqslant |F| \|q^{\eta}_{E,t} - q^{\delta}_{E,t}\|_2^2
\end{aligned}$$

where

- the first inequality holds by the Cauchy-Schwarz inequality,
- the second inequality holds because $\xi_{j,t}$ is a probability distribution over the actions set $[K]$,
- the last inequality holds because $|E| \leqslant |F|$

Therefore,

$$\|\pi^{\eta}_{E,t} - \pi^{\delta}_{E,t}\|_2 \leqslant \frac{|F|}{\varepsilon^2} \left(\varepsilon |\eta_T - \delta_T|(t-1) + \delta_T \sum_{s=1}^{t-1} \|\pi^{\eta}_{E,s} - \pi^{\delta}_{E,s}\|_2 \right).$$

Using the discrete Gronwall Lemma, for all $t \leqslant T_\varepsilon$,

$$\|\pi^{\eta}_{E,t} - \pi^{\delta}_{E,t}\|_2 \leqslant \frac{|F|}{\varepsilon} |\eta_T - \delta_T| T_\varepsilon \prod_{s=1}^{t-1} \left(1 + \frac{|F|}{\varepsilon^2} \delta_T \right) \leqslant \frac{|F|}{\varepsilon} |\eta_T - \delta_T| T_\varepsilon \exp\left(\frac{|F|}{\varepsilon^2} \delta_T T_\varepsilon \right).$$

If Assumption 3 is satisfied, then since $\delta_T \leqslant R_T = \frac{R}{\sqrt{T}}$ and $T_\varepsilon \leqslant \left(\frac{1}{|F|} - \frac{\varepsilon}{\rho} \right) \frac{\sqrt{T}}{R}$,

$$\|\pi^{\eta}_{E,t} - \pi^{\delta}_{E,t}\|_2 \leqslant \frac{|F|}{R\varepsilon} \left(\frac{1}{|F|} - \frac{\varepsilon}{\rho} \right) \exp\left(\frac{|F|}{\varepsilon^2} \left(\frac{1}{|F|} - \frac{\varepsilon}{\rho} \right) \right) |\eta - \delta|.$$

To conclude note that $x \to \ln(x)$ is $1/\varepsilon$-Lipschitz on $[\varepsilon, +\infty)$. $\square$