

EFFICIENT CONTEXTUAL PARTIAL MONITORING WITH GENERAL FUNCTION APPROXIMATION

Kapilan Balagopalan, Hao Qin

Department of Computer Science
University of Arizona
Tucson, AZ, USA
{kapilanbgp, hqin}@arizona.edu

Kwang-Sung Jun

Graduate School of Artificial Intelligence
POSTECH
Pohang, South Korea
kwangsungjun@gmail.com

Chicheng Zhang

Department of Computer Science
University of Arizona
Tucson, AZ, USA
chichengz@arizona.edu

ABSTRACT

We study finite stochastic partial monitoring (PM) games with side information, a paradigm of sequential learning with partial feedback. On each round $t \in [T]$, both player and environment observe a context and player chooses an action and environment simultaneously chooses an outcome. The player neither observes the outcome nor the loss; instead they only observe a feedback symbol which is partially informative about the outcome. We utilize the Estimation-to-Decisions (E2D) meta-algorithm of Foster et al. (2021) to design E2D.PMSIDE, a computationally-efficient framework for contextual partial monitoring, which uses an online regression oracle to efficiently estimate the outcome distributions based on historical (context, action, feedback symbol) tuples. Since the exploration rule given by the original E2D algorithm is intractable to compute, we propose two computationally-efficient exploration rules that achieve optimal regret orders in T : in the globally observable setting, our exploration rule is based on ϵ -greedy, while in the locally observable setting, our exploration rule uses inverse gap weighting (Foster & Rakhlin, 2020) which was previously only known to be effective in the bandit setting. We show that a variant of E2D.PMSIDE named OE2D.PMSIDE can be combined with offline regression oracles to give an efficient algorithm with optimal regret when the contexts are drawn independently from a fixed distribution. We evaluate our algorithm on a synthetic label-efficient prediction game and an LLM routing task, in which a learner decides from prompt, uncertainty and internal state features whether to escalate a query from a weak model to a strong one. We show that our method has superior performance than prior approaches.

1 INTRODUCTION

The contextual stochastic partial monitoring problem (Bartók & Szepesvári, 2012) is an online learning problem where the learner repeatedly takes actions so as to minimize its loss based on the environment’s partial feedback. At the beginning of the game, the learner and the environment are given a pair $\mathcal{G} = (\mathbf{L}, \Phi)$ of $k \times d$ matrices, where k is the number of actions available to the learner and d is the number of outcomes available to the environment. The *loss matrix* $\mathbf{L} \in [0, 1]^{k \times d}$ records the loss L_{ao} incurred when the learner plays action a and the environment’s outcome is o . The *feedback matrix* $\Phi \in \Sigma^{k \times d}$ records what the learner is shown instead: playing a against outcome o reveals the feedback symbol Φ_{ao} from a finite alphabet $\Sigma = \{1, \dots, m\}$. The interaction between the learner and the environment proceeds over T rounds. At round t the learner observes a context x_t (which may be chosen adversarially), samples an action $A_t \sim p_t(\cdot | x_t)$ according to a policy $p_t : \mathcal{X} \rightarrow \Delta(k)$, and the environment draws an outcome O_t from some stochastic distribution $q^*(x_t)$ independently of the learner’s randomization. The learner then observes the feedback symbol $\sigma_t = \Phi_{A_t O_t}$ and incurs the loss $L_{A_t O_t}$, which remains hidden. The goal of the learner is to minimize its regret, which is the

difference between its expected cumulative loss and that of the best policy in hindsight:

$$\text{Reg}_T := \sum_{t=1}^T \left(p_t^\top L q^*(x_t) - \min_{\lambda \in \Delta(k)} \lambda^\top L q^*(x_t) \right).$$

Contextual stochastic partial monitoring has many applications, such as learning from one-sided feedback (Helmbold et al., 2000; Sculley, 2007), dynamic pricing (Kleinberg & Leighton, 2003), data procurement in crowdsourcing (Singla et al., 2014), and label-efficient prediction (Helmbold et al., 1997). Below we give two examples:

Example 1: LLM routing. Modern Large language models (LLM)-based question-answering service optimizes its cost-efficiency by operating in a hybrid mode (Ong et al., 2025; Chen et al., 2023; An et al., 2026; Li et al., 2026) : it keeps two LLMs – one weak but cheap and one strong but expensive; and given a prompt, it judiciously decides which LLMs to route the prompt to, obtaining its response. At time step t , the service receives a new prompt represented by context x_t , and the weak model generates a response at negligible cost, and the router then either serves the weak response (a_1) or queries the strong model and serves its response (a_2). The environment’s (hidden) outcome is whether the weak response agrees with the strong one (o_1) or does not agree (o_2). We take the strong model to be correct and to cost 1 per query, and we incur cost of $\ell_{01} > 1$ when a served weak model’s response is wrong. Partial monitoring enters through the feedback: serving the weak response never reveals whether it was correct, so both outcomes yield the same symbol $\odot$, and only querying the strong model distinguishes them. The game’s loss matrix and feedback matrix are summarized below:

$$L = \begin{matrix} & o_1 & o_2 \\ \begin{matrix} a_1 \\ a_2 \end{matrix} & \begin{bmatrix} 0 & \ell_{01} \\ 1 & 1 \end{bmatrix} \end{matrix}, \quad \Phi = \begin{matrix} & o_1 & o_2 \\ \begin{matrix} a_1 \\ a_2 \end{matrix} & \begin{bmatrix} \odot & \odot \\ \perp & \wedge \end{bmatrix} \end{matrix}.$$

This game was also referred to as the Apple Tasting game Helmbold et al. (2000) in earlier works.

Example 2: contextual label efficient prediction game (Lattimore & Szepesvári, 2020; Helmbold et al., 1997). Suppose we would like to learn a spam filter online, which predicts whether an email is spam (o_1) or not (o_2) based on its text features (x_t) and directs it to the inbox (a_2) or spam folder (a_3). There is also a third option (a_1) where the system can ask the recipient/expert for classification, which is accurate but incurs a fixed cost. Only when the learner chooses action a_1 does it get to see the true label of the email. The game’s loss matrix and feedback matrix are summarized below:

$$L = \begin{matrix} & o_1 & o_2 \\ \begin{matrix} a_1 \\ a_2 \\ a_3 \end{matrix} & \begin{bmatrix} c & c \\ 1 & 0 \\ 0 & 1 \end{bmatrix} \end{matrix}, \quad \Phi = \begin{matrix} & o_1 & o_2 \\ \begin{matrix} a_1 \\ a_2 \\ a_3 \end{matrix} & \begin{bmatrix} \perp & \wedge \\ \odot & \odot \\ \odot & \odot \end{bmatrix} \end{matrix}.$$

Depending on the feedback structure, finite stochastic partial-monitoring games are classified into four categories (classification theorem): trivial, easy, hard, and hopeless games (Bartók et al., 2011; Lattimore & Szepesvári, 2020). In hopeless games, the feedback is too uninformative to support meaningful learning, and the regret might grow linearly with the time horizon ($\Omega(T)$). In contrast, in trivial games, there exists an action which is optimal for every outcome; consequently, they can be solved with zero regret. The main cases requiring more refined analysis are easy ($\Theta(\sqrt{T})$) and hard games ($\Theta(T^{2/3})$). Both classes satisfy global observability, which we define shortly. In addition, easy games satisfy local observability, whereas hard games do not.

Related work: partial monitoring with side information. We adopt the contextual partial monitoring setting of Bartók et al. (2012a); Heuillet et al. (2024), and use the terms *contextual partial monitoring* and *partial monitoring with side information* interchangeably for it. The classification of easy and hard games established for finite stochastic non-contextual partial monitoring (Bartók et al., 2011; Lattimore & Szepesvári, 2020) extends to this setting, at least in the minimax regret’s dependence on T . First, the non-contextual lower bounds of Bartók et al. (2011) transfer immediately, since a non-contextual game can be viewed as a contextual game with a dummy context. Second, for the upper bounds, Foster et al. (2022) shows the equivalence between the Decision-Estimation Coefficient Foster et al. (2021) and the exploration-by-optimization (ExO) objective of Lattimore & Szepesvári (2020), which implies that the contextual version of the Estimation-to-Decision algorithm (Foster et al., 2021, Section 8) achieve regrets that match optimal rates given by ExO. These optimal statistical guarantees come at a computational cost, however: the algorithm attaining them selects each action distribution

by solving a minimax problem at every round, which is not known to be tractable (see Section 2.3 for further discussion). Furthermore, our work is closely related to Lattimore & Szepesvári (2019; 2020); Lattimore (2022); Lattimore & Gyorgy (2021) and Foster et al. (2022). A detailed discussion is deferred to Appendix A.

Meanwhile, computationally efficient algorithms have been proposed for contextual stochastic partial monitoring, by imposing strong assumptions on the function class that models the true outcome distribution q^* . Utilizing generalized linear models for modeling (functions of) q^* , Bartók et al. (2012a) proposed CBPSIDE, a deterministic algorithm based on confidence intervals that extends UCB-style methods from the bandit setting; it applies only to easy games, where it achieves $O(\sqrt{T})$ regret. Heuillet et al. (2024) later introduced CBPSIDE* and its randomized counterpart RANDCBPSIDE*, which cover both classes, attaining $O(\sqrt{T})$ regret on easy games and $O(T^{2/3})$ on hard ones. Kirschner et al. (2023) studies the linear partial monitoring and its contextual extension, which generalizes contextual finite partial monitoring; however, they require that the mappings from (context, action) to the expected reward and the signal feedback to be (context-independent) linear functions.

This motivates the question that drives our work: are there computationally efficient algorithms for contextual partial monitoring that support general function approximation and attain the optimal rates? In this paper, we answer this question affirmatively. Specifically, we make the following contributions:

Contributions. (1) we propose an algorithmic framework E2D.PMSIDE that reduces stochastic contextual partial monitoring to online regression. It is based on the E2D (Estimation to Decision) framework (Foster et al., 2021), where the algorithm aims at finding an exploration distribution that certifies a low Decision-Estimation Coefficient (See Section 2.3 for its definition); our key observation is that this can be done by optimizing a surrogate objective that trades off between greediness and coverage, which paves the way for the design of efficient algorithms for contextual partial monitoring. (2) Based on the above relaxed objective, we instantiate E2D.PMSIDE in the globally observable and locally observable games to obtain ϵ -GREEDY.PMSIDE and SQUARECB.PMSIDE algorithms respectively, achieving regret of optimal order in T in their respective settings. These, to our knowledge, are the first *computationally efficient* contextual partial monitoring algorithms with *general function approximation*. (3) When the contexts are iid, we propose a variant of E2D.PMSIDE named OE2D.PMSIDE, which is an efficient stochastic contextual partial monitoring algorithm based on offline regressions, a computational primitive that can admit simpler implementations in practice. (4) We evaluate SQUARECB.PMSIDE in two contextual partial monitoring tasks: label-efficient prediction and LLM routing, and show that it outperforms several baselines.

2 PRELIMINARIES

Basic Notation. We write $\mathbf{0}$ and $\mathbf{1}$ for the all-zeros and all-ones column vectors, $e_1, \dots, e_k$ for the standard basis vectors, and bold symbols for matrices. For a vector or matrix A , $\|A\|_\infty = \max_{ij} |A_{ij}|$ is the entrywise maximum norm, M_i denotes the vector generated by i -th row, and M_{ij} denotes the element at i -th row and j -th column. For two distributions u and v , $D_H^2(u|v)$ denotes their squared Hellinger distance. Let $\text{co}(\mathcal{S})$ denote the convex hull of a set $\mathcal{S}$. For a subset $\mathcal{A} \subseteq [k]$ we write $\Delta(\mathcal{A}) := \{\lambda \in \Delta(k) : \lambda_b = 0 \text{ for all } b \notin \mathcal{A}\}$ for the face of $\Delta(k)$ supported on $\mathcal{A}$. Elements of $\Delta(\mathcal{A})$ remain k -dimensional vectors; they simply place no mass outside $\mathcal{A}$.

Signal matrix. (Komiyama et al. (2015)) Given a feedback matrix Φ , it is convenient to describe the feedback of a single action as a linear map. The *signal matrix* of action a is the $m \times d$ matrix S_a with $S_a[\sigma, o] = \mathbb{1}\{\Phi_{ao} = \sigma\}$. Writing $\Delta(n)$ for the probability simplex on n elements, if the outcome is drawn from $q \in \Delta(d)$ then $S_a q \in \Delta(m)$ is precisely the distribution of the symbol observed on playing a . Two candidate outcome distributions q_1, q_2 are therefore distinguishable through action a exactly to the extent that $S_a q_1$ and $S_a q_2$ differ; and it is this quantity, rather than the gap between q_1 and q_2 themselves, that governs what the learner can hope to identify. Recall that in our stochastic contextual partial monitoring problem, each round is preceded by a context x_t drawn from a set $\mathcal{X}$, and the outcome distribution depends on it: outcomes follow $q^*(x_t)$ for an unknown $q^* : \mathcal{X} \rightarrow \Delta(d)$. The learner is given a model class $\mathcal{Q} \subseteq \{q : \mathcal{X} \rightarrow \Delta(d)\}$, which we assume is well-specified.

Assumption 2.1 (Realizability). The model class $\mathcal{Q}$ contains the true outcome distribution q^* .

2.1 STRUCTURE OF PARTIAL MONITORING GAMES

We now review some basic structures of partial monitoring games revealed in prior works.

Definition 2.1 (Cell decomposition, Bartók et al. (2012b)). The *cell* $\mathcal{C}_a$ of action a is defined as the subset of outcome distributions where action a is optimal. Formally, $\mathcal{C}_a = \{q \in \Delta(d) : (e_a - e_b)^\top Lq \leq 0, \forall b \in [k]\}$. The collection $\{\mathcal{C}_a : a \in [k]\}$ is called the cell decomposition.

Neighborhood graph (Lattimore & Szepesvári, 2020). Cells that meet in a face of dimension $d - 2$ correspond to actions that can be optimal simultaneously, and this adjacency structure organizes the analysis that follows. Call an action *Pareto optimal* when its cell is full dimensional, $\dim(\mathcal{C}_a) = d - 1$, taking the dimension of a polytope to be that of its affine hull, and write Π for the set of such actions together with $\Delta(\Pi) := \{\lambda \in \Delta(k) : \lambda_b = 0 \text{ for all } b \notin \Pi\}$ for the corresponding face of $\Delta(k)$. The remaining actions are of two kinds: a is *degenerate* when $\mathcal{C}_a \neq \emptyset$ but $\dim(\mathcal{C}_a) \leq d - 2$, and *dominated* when $\mathcal{C}_a = \emptyset$. Since $\dim(\Delta(d)) = d - 1$, neither kind is ever uniquely optimal; each is optimal only on a set of measure zero. A game with no degenerate actions is called *non-degenerate*, and two actions are *duplicates* when their loss rows coincide, $(e_a - e_b)^\top L = \mathbf{0}$. Two Pareto optimal actions are *neighbors* when their cells meet in a face of the largest dimension short of coincidence, $\dim(\mathcal{C}_a \cap \mathcal{C}_b) = d - 2$; the intersection reaches dimension $d - 1$ only for duplicates. Writing $\mathcal{E}$ for the resulting edge set, $(\Pi, \mathcal{E})$ is the *neighborhood graph* of $\mathcal{G}$.

2.2 OBSERVABILITY

In partial-monitoring games, the learner observes feedback rather than the losses or rewards associated with its actions. Therefore, a central challenge is to infer (differences of) losses of actions from the observed feedback symbol. Whether this is possible depends on the interplay between the loss matrix and the feedback matrix.

Definition 2.2 (Global observability, adapted from Bartók et al. (2012b) and Lattimore & Szepesvári (2020)). A partial monitoring game is *globally observable* if, for every pair of actions $(a, b) \in [k] \times [k]$, there exists a *witness function* $w_{(a,b)} : [k] \times \Sigma \rightarrow \mathbb{R}$ such that

$$L_{ao} - L_{bo} = \sum_{c=1}^k w_{(a,b)}(c, \Phi_{co}), \quad \text{for all } o \in [d].$$

Definition 2.3 (Local observability, Lattimore & Szepesvári (2020)). A globally observable, *non-degenerate* partial monitoring game is *locally observable* if for every edge $e = (a, b) \in \mathcal{E}$ of the neighborhood graph the witness function w_e of Definition 2.2 may be chosen to be supported only on the endpoints, that is $w_e(c, \cdot) \equiv 0$ for all $c \notin \{a, b\}$, so that

$$L_{ao} - L_{bo} = \sum_{c \in \{a,b\}} w_e(c, \Phi_{co}), \quad \text{for all } o \in [d].$$

By definition, every locally observable game is a globally observable game. Under global observability, the loss difference $L_{ao} - L_{bo}$ may depend on feedback from all actions $c \in [k]$, whereas under local observability, it can be expressed using only the feedback from the sequence of actions in $\mathcal{E}$ connecting a and b . Consequently, locally observable games permit more direct loss-difference estimation than globally observable games in general. Since trivial games and hopeless games are not interesting from an algorithm design perspective, for the remainder of the paper we focus on globally observable games.

Assumption 2.2 (Global observability and non-degeneracy; Lattimore & Szepesvári (2020)). We assume that the partial-monitoring game $\mathcal{G} = (L, \Phi)$ is globally observable, non-degenerate, and contains no duplicate actions.

For simplicity, we maintain this assumption throughout the remainder of the analysis, following Lattimore & Szepesvári (2020). We discuss implications of non-degeneracy and possible relaxations in Appendix C.4.

2.3 THE E2D META-ALGORITHM FOR INTERACTIVE DECISION MAKING

Foster et al. (2021) proposes E2D (Estimation-to-Decision), a meta-algorithm for interactive decision making, which reduces interactive decision making to online estimation. The E2D meta-algorithm yielded many practical algorithms for decision making, especially for contextual bandit problems (e.g., Foster & Rakhlin, 2020; Zhu et al., 2022; Zhu & Mineiro, 2022; Zhang et al., 2023). In a nutshell, at each round, it calls an online model estimation oracle to obtain an estimated model $\hat{q}$; based on $\hat{q}$ and context x , it solves a minimax optimization problem to obtain an exploration distribution p to take

actions. Specialized to partial monitoring, the optimization problem can be written as:

$$\min_{p \in \Delta(k)} \sup_{q \in \mathcal{Q}_x, \lambda \in \Delta(k)} \underbrace{\left[p^\top \mathbf{L}q(x) - \lambda^\top \mathbf{L}q(x) - \gamma \cdot \mathbb{E}_{a \sim p} \left[D_H^2(\mathbf{S}_a \hat{q}(x) \mid \mathbf{S}_a q(x)) \right] \right]}_{\mathcal{V}_\gamma^{\text{dec}}(\hat{q}(x), \mathcal{Q}_x, p)}, \quad (1)$$

where $\mathcal{Q}_x := \{q(x) : q \in \mathcal{Q}\} \subseteq \Delta(d)$ denotes the model class $\mathcal{Q}$ restricted to x . The objective explicitly trades off between regret and information gain: the first part $p^\top \mathbf{L}q(x) - \lambda^\top \mathbf{L}q(x)$ encourages p to have a low regret, while the second part $-\gamma \cdot \mathbb{E}_{a \sim p} \left[D_H^2(\mathbf{S}_a \hat{q}(x) \mid \mathbf{S}_a q(x)) \right]$ encourages p to explore diverse actions to correct $\hat{q}$ and gain new information about q^* . The minimax value of the objective is called the Decision-Estimation Coefficient (DEC) (Foster et al., 2021), denoted as $\text{dec}_\gamma(\hat{q}(x), \mathcal{Q}_x)$. To help the learner make decisions efficiently, E2D requires access to an oracle that can provide accurate estimation of the outcome distributions for all contexts online, which, when specialized to our stochastic contextual partial monitoring setting, is defined as:

Definition 2.4. An *online outcome estimation oracle* Alg_{On} is an algorithm that estimates the unknown outcome distribution q^* from the data in a sequential fashion. That is, at each round $t \in [T]$, given the feedback symbols collected so far, the oracle produces an estimate

$$\hat{q}_t = \text{Alg}_{\text{On}}^{(t)} \left(\left\{ (A_s, x_s, \sigma_s) \right\}_{s=1}^{t-1} \right) \in \text{co}(\mathcal{Q}),$$

where $\text{Alg}_{\text{On}}^{(t)}$ denotes the map applied at round t and $\left\{ (A_s, x_s, \sigma_s) \right\}_{s=1}^{t-1}$ is the history available at the beginning of round t (note that σ_s is an observed feedback symbol, but not the incurred loss).

Alg_{On} is said to satisfy a *Hellinger regret bound* of $\overline{\text{Reg}}_{\text{On}}^H(\mathcal{Q}, T, \delta)$ if its online Hellinger regret

$$\text{Reg}_{\text{On}}^H(\mathcal{Q}, T) := \sum_{t=1}^T \mathbb{E}_{a \sim p_t(\cdot \mid x_t)} \left[D_H^2(\mathbf{S}_a \hat{q}_t(x_t) \mid \mathbf{S}_a q^*(x_t)) \right], \quad (2)$$

is at most $\overline{\text{Reg}}_{\text{On}}^H(\mathcal{Q}, T, \delta)$ with probability $1 - \delta$.

We note that instead of requiring $\hat{q}_t(x_t)$ to be pointwise close to $q^*(x_t)$ on average, Definition 2.4 only requires their associated feedback distributions to be close on average. This is reasonable since we can only use the feedback σ_s 's to guide our online learning of $\hat{q}_t$; in the extreme case, it is possible that different q 's result in the same feedback distribution, making the underlying q unidentifiable. Appendix B gives two concrete examples of the online outcome estimation oracle. For example, For finite $\mathcal{Q}$, Vovk's aggregating algorithm (Vovk, 1995) attains $\overline{\text{Reg}}_{\text{On}}^H(\mathcal{Q}, T, \delta) = O(\log(|\mathcal{Q}|/\delta))$. Equipped with an online outcome estimation oracle, E2D is known to have a regret bound of

$$\text{Reg}_T \leq O \left(T \max_{x \in \mathcal{X}} \text{dec}_\gamma(\hat{q}(x), \mathcal{Q}_x) + \gamma \overline{\text{Reg}}_{\text{On}}^H(\mathcal{Q}, T, \delta) \right),$$

which translates to $O(T^{2/3})$ and $O(\sqrt{T})$ regrets with optimal tuning of γ in globally and locally observable settings, respectively (Foster et al., 2021). However, a key drawback of the naive E2D algorithm is that directly solving the minimax optimization problem (Equation (1)) is computationally intractable: although the objective is linear in p , it is not jointly concave in (q, λ) . Nevertheless, it was shown in Foster et al. (2021) that as long as distribution p certifies a small value of $\text{dec}_\gamma(\hat{q}(x), \mathcal{Q}_x)$ (Equation (1)), i.e., it achieves a small $\mathcal{V}_\gamma^{\text{dec}}(\hat{q}(x), \mathcal{Q}_x, p)$ value, E2D will have small regret. As we will see, our algorithm design will utilize this observation to address the computational challenge.

3 EFFICIENT ALGORITHM FOR CONTEXTUAL PARTIAL MONITORING

In this section we present E2D.PMSIDE (Estimation-to-Decision for Partial Monitoring with Side information, Algorithm 1), which can be viewed as a specialization of the E2D meta-algorithm. At every round t , E2D.PMSIDE obtains an outcome distribution estimate $\hat{q}_t(\cdot)$ (line 4) from an online outcome estimation oracle with Hellinger regret guarantee. Then it uses $\hat{q}_t(\cdot)$ and observed context x_t to construct an exploration distribution p_t that approximately minimizes an objective function $\mathcal{V}_\gamma(\hat{q}_t(x_t), p)$, a relaxation of the DEC objective $\mathcal{V}_\gamma^{\text{dec}}(\hat{q}_t(x_t), \mathcal{Q}_{x_t}, p)$ (line 5). The choice of p_t balances two competing demands. The first term in the objective measures the loss of p_t relative to a comparator policy λ under the estimated outcome distribution $\hat{q}_t(x_t)$, and is small when p_t exploits what has been learned. The second term is a *coverage* term (which we will define shortly) that measures the amenability of using feedback from following p in evaluating λ 's loss; the smaller the

Algorithm 1 E2D.PMSIDE1: **parameters:**Online estimation oracle Alg_{On} Exploration parameter $\gamma > 0$ 2: **for** $t = 1, 2, \dots, T$ **do**3: Receive context x_t 4: Compute estimate $\hat{q}_t = \text{Alg}_{\text{On}}^{(t)}(\{A_s, x_s, \sigma_s\}_{s=1}^{t-1})$ 5: Find p_t that approximately minimizes

$$\mathcal{V}_\gamma(\hat{q}_t(x_t), p) := \sup_{\lambda \in \Delta(\Pi)} \underbrace{p^\top \mathbf{L} \hat{q}_t(x_t) - \lambda^\top \mathbf{L} \hat{q}_t(x_t)}_{\text{exploit}} + \frac{1}{\gamma} \cdot \underbrace{\text{Coverage}(p, \lambda)}_{\text{explore}}.$$

using ϵ -GREEDY.PMSIDE (Alg. 2) or SQUARECB.PMSIDE (Alg. 3) for globally and locally observable settings, respectively.

6: Choose action $A_t \sim p_t$ and update the estimation oracle with (A_t, x_t, σ_t) 7: **end for**

coverage value the better p is for exploration. The exploration parameter γ fixes the rate of exchange between the two. We now define coverage formally.

Definition 3.1 (Coverage). We define the coverage of action distribution $p \in \Delta(k)$ for action distribution $\lambda \in \Delta(k)$ for partial monitoring game $\mathcal{G} = (\mathbf{L}, \Phi)$ as

$$\text{Coverage}(p, \lambda) := \sup_{q_1, q_2 \in \Delta(d)} \frac{\left((\lambda - p)^\top \mathbf{L} (q_1 - q_2) \right)^2}{\mathbb{E}_{a \sim p} \left[D_H^2(\mathbf{S}_a q_1 \mid \mathbf{S}_a q_2) \right]}. \quad (\text{with convention } \frac{0}{0} = 0)$$

To make sense of this definition it may be helpful to think about the following hypothetical offline policy evaluation problem in non-contextual stochastic partial monitoring: suppose the outcome distribution q^* is unknown; we use a behavior policy p to take action and collect feedback symbols, so as to estimate the expected loss of λ (more precisely, the expected loss difference between λ and p); how much utility does each data point carry for such evaluation? Let q_1 be q^* and q_2 be $\hat{q}$, our estimate of q^* based on data. The numerator thus measures the squared policy evaluation error; the denominator measures the in-distribution estimation error, which we expect to decrease with more samples collected. A small $\text{Coverage}(p, \lambda)$ implies that feedback signals via following p is useful in estimating p and λ 's loss difference.

Our coverage in the above definition bears some similarities with the coverage defined in the contextual bandit setting by Qin & Zhang (2026b), but is different in two important ways: (1) here, the numerator and the denominator are with respect to the loss matrix (evaluation error) and the signal matrix (information gain), respectively, where in the bandit setting they are defined both using the loss function; (2) it is well-known that for some globally observable partial monitoring problems the absolute value of a loss of an action is not identifiable (Lattimore & Szepesvári, 2020, Example 37.9), and only loss differences between action pairs is identifiable; thus, the numerator in the coverage measures the difference of losses of two policies. Our coverage also shares some similarities with global and local alignment constants of (Kirschner et al., 2023, Eqs. (15), (16)), whose numerator $(\langle \phi_a - \phi_b, v \rangle)^2$ can be interpreted as some estimation error of the loss differences between two actions a, b , and the denominator $\|M_c v\|^2$ can be interpreted as the information gain of taking action c in correcting the model estimation error v . Instead of measuring coverage between an action and a pair of actions, our definition measures the coverage between two action distributions.

We now present the regret guarantee of E2D.PMSIDE:

Theorem 3.1 (Regret of E2D.PMSIDE). Suppose Alg_{On} has a Hellinger regret bound $\overline{\text{Reg}}_{\text{On}}^{\text{H}}(\mathcal{Q}, T, \delta)$. Fix $\gamma > 0$. Then, for any $\delta \in (0, 1)$, with probability at least $1 - \delta$,

$$\text{Reg}_T \leq T \cdot \max_{t=1}^T \mathcal{V}_\gamma(\hat{q}_t(x_t), p_t) + \gamma \cdot \overline{\text{Reg}}_{\text{On}}^{\text{H}}(\mathcal{Q}, T, \delta).$$

Its proof is deferred to Appendix C.1. In a nutshell, our theorem follows from the meta-regret theorem of Foster et al. (2021): as long as all action distributions p_t certify that $\text{dec}_\gamma(\hat{q}(x), \mathcal{Q}_x)$ has an upper bound, E2D has a small regret in terms of that bound. Indeed, our p_t here certifies that

Algorithm 2 ϵ -GREEDY.PMSIDE for globally observable games

-
- 1: **parameters:**
 - 2: Context x with predicted outcome distribution $\hat{q}(x)$
 - 3: Exploration parameter $\gamma > 16k^2\beta^2$
 - 4: Let $\hat{b} \in \arg \min_{a \in \Pi} e_a^\top \mathbf{L} \hat{q}(x)$
 - 5: **return** $p = (1 - \epsilon)e_{\hat{b}} + \epsilon \frac{1}{k} \mathbf{1}$ for $\epsilon = k\beta/\sqrt{\gamma}$
-

$\text{dec}_\gamma(\hat{q}(x), \mathcal{Q}_x)$ is upper bounded by $\mathcal{V}_\gamma(\hat{q}_t(x_t), p_t)$ (Please see Remark 3) ; and ideally, we should choose exploration distribution p_t 's that make $\mathcal{V}_\gamma(\hat{q}_t(x_t), p_t)$'s as small as possible.

Unfortunately, finding p_t that exactly minimizes $\mathcal{V}_\gamma(\hat{q}_t(x_t), p)$ may still be computationally intractable; to see this, note that the definition of coverage involves supremum over a pair of outcome distributions q_1, q_2 , which is not a convex optimization problem. This motivates the question: can we find computationally efficient exploration rules $x \mapsto p(x)$ that *approximately minimize* $\mathcal{V}_\gamma(\hat{q}_t(x), p(x))$ for all x ? Indeed, we show in the following two subsections that such efficient exploration rule exist and yield efficient implementations of E2D.PMSIDE.

3.1 EFFICIENT ALGORITHM FOR GLOBALLY OBSERVABLE GAMES

We show that in the globally observable setting ϵ -GREEDY.PMSIDE (Alg. 2), yields a policy p satisfying $\mathcal{V}_\gamma(\hat{q}(x), p(x)) \leq O(\sqrt{1/\gamma})$, which gives an $O(T^{2/3})$ regret bound when plugged into E2D.PMSIDE. ϵ -GREEDY.PMSIDE identifies the empirical best action $\hat{b}$ (Line 4) and assigns it probability $1 - \epsilon$, spreading the remaining ϵ uniformly over all k actions (Line 5).

Lemma 3.1 (Guarantee for Alg. 2 for globally observable games). Let $\mathcal{G} = (\mathbf{L}, \Phi)$ be a globally observable partial monitoring game in the sense of Definition 2.2, and set $\beta := \max_{(a,b) \in [k] \times [k]} \|w_{(a,b)}\|_\infty$. We have for any $\hat{q} \in \Delta(d)$,

$$\mathcal{V}_\gamma(\hat{q}, p) \leq \max_{\lambda \in \Delta(\Pi)} p^\top \mathbf{L} \hat{q} - \lambda^\top \mathbf{L} \hat{q} + \frac{16\beta^2}{\gamma} \sum_{c=1}^k \frac{1}{p_c}.$$

Thus, for every $\gamma > 16k^2\beta^2$, every $\hat{q} \in \Delta(d)$, Alg. 2 outputs p such that

$$\mathcal{V}_\gamma(\hat{q}, p) \leq O\left(k\beta\sqrt{\frac{1}{\gamma}}\right).$$

Proof is deferred to Appendix C.2. Furthermore, by combining Theorem 3.1 and Lemma 3.1, our algorithm achieves the following minimax rate for globally observable games.

Corollary 1 (Regret of E2D.PMSIDE for globally observable games). Suppose Alg_{On} has a Hellinger regret bound $\overline{\text{Reg}}_{\text{On}}^{\text{H}}(\mathcal{Q}, T, \delta)$. For a globally observable partial monitoring game $(\mathbf{L}, \Phi)$, set $\beta = \max_{(a,b) \in [k] \times [k]} \|w_{(a,b)}\|_\infty$. Then, for any $\delta \in (0, 1)$, and for any large enough T that satisfies $\gamma > 16k^2\beta^2$ with $\gamma = (Tk\beta / \overline{\text{Reg}}_{\text{On}}^{\text{H}}(\mathcal{Q}, T, \delta))^{2/3}$, with probability of at least $1 - \delta$, E2D.PMSIDE run with Alg 2 as the exploration rule satisfies,

$$\text{Reg}_T \leq O\left(\sqrt[3]{T^2 k^2 \beta^2 \overline{\text{Reg}}_{\text{On}}^{\text{H}}(\mathcal{Q}, T, \delta)}\right).$$

In the setting when $\mathcal{Q}$ is finite, setting the estimation oracle to be Vovk's aggregating algorithm, the above regret bound is of order $O(\sqrt[3]{T^2 k^2 \beta^2 \log |\mathcal{Q}|})$. In the noncontextual setting, when $\mathcal{Q} = \Delta(d_{\mathcal{Q}})$, we can set the estimation oracle to be any efficient exp-concave optimization algorithm (Hazan et al., 2007), which has $\overline{\text{Reg}}_{\text{On}}^{\text{H}}(\mathcal{Q}, T, \delta) = O(d_{\mathcal{Q}} \ln T)$, implying E2D.PMSIDE has a regret of $\tilde{O}(\sqrt[3]{T^2 k^2 \beta^2 d_{\mathcal{Q}}})$. This matches the lower bound of non-contextual stochastic partial monitoring (Bartók et al., 2011) with a simple computationally efficient approach.

3.2 EFFICIENT ALGORITHM FOR LOCALLY OBSERVABLE GAMES

In the locally observable setting, we aim to give an efficient exploration rule that finds p such that $\mathcal{V}_\gamma(\hat{q}_t(x), p(x)) \leq O(\frac{1}{\gamma})$, which yields a $O(\sqrt{T})$ regret when plugged into E2D.PMSIDE. Under

global observability, to estimate the loss difference between two actions (a, b) , its associated witness function $w_{(a,b)}$ may be supported on all actions (Definition 2.2), so recovering the difference between a single pair of actions can require playing all k of them; this is why the certifying policy of Algorithm 2 mixes with the uniform distribution over $[k]$. In contrast, under local observability, the loss difference between two *neighboring* Pareto optimal actions can be expressed using the feedback of those two actions alone (Definition 2.3), and the difference between an arbitrary pair of Pareto optimal actions is then recovered by chaining such local relations along a path joining them in the neighborhood graph, so that only the actions encountered along that path need be played. Perhaps surprisingly, we show that Inverse Gap Weighting, an exploration rule that was previously known to be effective for exploration in contextual bandits (SQUARECB.PMSIDE, Alg. 3), yields an optimal-regret algorithm here without modification.

Lemma 3.2 (Guarantee for Alg. 3 for locally observable games). Let $\mathcal{G} = (\mathbf{L}, \Phi)$ be a locally observable partial monitoring game in the sense of Definition 2.3. For every $\gamma > 0$, and every $\hat{q} \in \Delta(d)$, Alg. 3 run with $\kappa = \frac{\gamma}{8km^2}$ and $\mu = k$, outputs p such that

$$\mathcal{V}_\gamma(\hat{q}, p) \leq O\left(\frac{m^2 k^2}{\gamma}\right).$$

Proof is deferred to Appendix C.3. Furthermore, by combining Theorem 3.1 and Lemma 3.2, we can achieve the following minimax rate for locally observable games.

Corollary 2 (Regret of E2D.PMSIDE for locally observable games). Suppose Alg_{On} has a Hellinger regret bound $\overline{\text{Reg}}_{\text{On}}^{\text{H}}(\mathcal{Q}, T, \delta)$. Let Assumption 2.2 hold, and suppose in addition that the game is locally observable. Then, for any $\delta \in (0, 1)$, with probability of at least $1 - \delta$, E2D.PMSIDE run with Alg. 3 as the exploration rule (with $\kappa = \frac{\gamma}{8km^2}$ and $\mu = k$), with $\gamma = (Tm^2k^2 / \overline{\text{Reg}}_{\text{On}}^{\text{H}}(\mathcal{Q}, T, \delta))^{1/2}$ satisfies

$$\text{Reg}_T \leq O\left(\sqrt{T \cdot m^2 k^2 \cdot \overline{\text{Reg}}_{\text{On}}^{\text{H}}(\mathcal{Q}, T, \delta)}\right).$$

For finite $\mathcal{Q}$, Vovk’s aggregating algorithm gives $O(\sqrt{Tm^2k^2 \log|\mathcal{Q}|})$; in the non-contextual setting with $\mathcal{Q} = \Delta(d_{\mathcal{Q}})$, an exp-concave oracle (Hazan et al., 2007) gives $\tilde{O}(\sqrt{Tm^2k^2 d_{\mathcal{Q}}})$. Both match the minimax rate in T for this class. We next explain the key idea underlying the proof of Lemma 3.2, which hinges on a sparse relaxation of the objective $\mathcal{V}_\gamma(\hat{q}, p)$.

3.2.1 SPARSE RELAXATION OF $\mathcal{V}_\gamma(\hat{q}, p)$

We aim to find tractable relaxations of $\mathcal{V}_\gamma$ which sheds lights on finding efficient exploration strategies for approximately minimizing $\mathcal{V}_\gamma$. In particular, we aim to exploit the structured feedback mechanism of locally observable games to restrict exploratory actions to those lying along paths that connect pairs of actions whose loss differences are of interest. The neighborhood graph $(\Pi, \mathcal{E})$ is connected. Assign an arbitrary action $\text{root}_{\mathcal{T}}$ as the root and choose a spanning in-tree $\mathcal{T}$ (See Figure 1) such that, there is a unique directed path $\text{path}_{\mathcal{T}}(a)$ from any action $a \in \Pi$ to $\text{root}_{\mathcal{T}} \in \Pi$. Now, we can relax the $\mathcal{V}_\gamma(\hat{q}, p)$ as follows,

Lemma 3.3. Let $\mathcal{G} = (\mathbf{L}, \Phi)$ be a locally observable partial monitoring game in the sense of Definition 2.3. Let $\mathcal{T}$ be an in-tree of the neighborhood graph. For all $p \in \Delta(\Pi)$, we have

$$\mathcal{V}_\gamma(\hat{q}(x), p) \leq \max_{\lambda \in \Delta(\Pi)} p^T \mathbf{L} \hat{q}(x) - \lambda^T \mathbf{L} \hat{q}(x) + \frac{32}{\gamma} \max_{e \in \mathcal{E}} \|w_e\|_\infty^2 \sum_{(b \in \Pi, c \in \Pi): c \in \text{path}_{\mathcal{T}}(b)} \left(\frac{\lambda_b}{p_c} + \frac{p_b}{p_c} \right). \quad (3)$$

Proof is deferred to Appendix C.3. Above result is true for any in-tree with an arbitrary root. However, in general, we do not want the paths to be arbitrary: for a “bad” action c and a “good” action b (i.e., $e_c^\top \mathbf{L} \hat{q}$ is large while $e_b^\top \mathbf{L} \hat{q}$ is small), we would like p_c to be small and p_b to be large to ensure $p^T \mathbf{L} \hat{q}(x) - \lambda^T \mathbf{L} \hat{q}(x)$ to be small; however, if c happens to lie in $\text{path}_{\mathcal{T}}(b)$, that can cause the $\frac{\lambda_b}{p_c}$ term to be very large. How should the in-tree $\mathcal{T}$ be chosen so that the actions lying on each path are ones we are willing to play? Our answer is to let the tree follow the estimated losses. We take $\mathcal{T}$ to be $\mathbf{L} \hat{q}(x)$ -decreasing (that is for a directed edge $e = (b, c)$, $e_b^\top \mathbf{L} \hat{q}(x) \geq e_c^\top \mathbf{L} \hat{q}(x)$; See Def C.1 for formal definition), so that estimated loss decreases along every path toward the root and, in particular, the

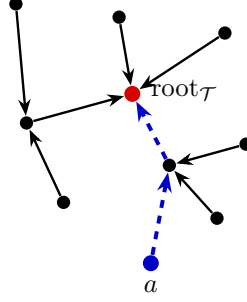


Figure 1: A spanning in-tree $\mathcal{T}$ of the neighborhood graph $(\Pi, \mathcal{E})$, with root $\text{root}_{\mathcal{T}}$; every edge is directed toward the root. The dashed edges form $\text{path}_{\mathcal{T}}(a)$, the directed path from action a to the root. The construction follows Lattimore & Szepesvári (2020).

Algorithm 3 SQUARECB.PMSIDE for Locally observable games

- 1: **parameters:**
 - 2: Context x with predicted outcome distribution $\hat{q}(x)$
 - 3: Exploration parameter $\gamma > 0$, $\mu = k$ and $\kappa = \frac{\gamma}{8km^2}$
 - 4: For each action $a \in \Pi$, compute $\hat{\ell}_a = e_a^\top \mathbf{L} \hat{q}(x)$ and let $b = \arg \min_{a \in \Pi} \hat{\ell}_{t,a}$
 - 5: For each $a \in \Pi \setminus \{b\}$, define $p_a = \frac{1}{\mu + \kappa(\hat{\ell}_a - \hat{\ell}_b)}$; let $p_b = 1 - \sum_{a \in \Pi \setminus \{b\}} p_a$.
 - 6: **return** p
-

empirical best action sits at the root. With this choice of $\mathcal{T}$, and the observation that p according to SQUARECB.PMSIDE is monotonically decreasing of in the loss, the coverage term is upper bounded by $k \sum_b (\frac{\lambda_b}{p_b} + 1)$, which allows us to bound $\mathcal{V}_\gamma(q, p)$ subsequently.

Offline oracle-efficient algorithms In Appendix D, we present a close variant of E2D.PMSIDE, namely OE2D.PMSIDE (Offline Estimation to Decisions for Partial Monitoring with Side Information), that efficiently reduces contextual partial monitoring to *offline* estimation. Our OE2D.PMSIDE generalizes FALCON (Simchi-Levi & Xu, 2022) and its linear version (Xu & Zeevi, 2020). Similar to E2D.PMSIDE, OE2D.PMSIDE has two versions, one that employs Alg. 2 for the globally observable setting, the other employs Alg. 3 for the locally observable setting, which enjoys $O(T^{2/3})$ and $O(\sqrt{T})$ regret in the respective settings. Compared with E2D.PMSIDE that makes $O(T)$ calls to the online outcome estimation oracle, OE2D.PMSIDE only makes $O(\log T)$ calls to the offline outcome estimation oracle, making it suitable for applications with low computational cost requirements.

4 EMPIRICAL STUDIES

We evaluate SQUARECB.PMSIDE in two settings we discussed in Section 1 (Examples 1 & 2): LLM routing and contextual label-efficient prediction game.

LLM routing. We evaluate the routing problem on a cached version of the BoolQ benchmark containing 12,648 valid responses from both models. We use Qwen2.5-0.5B-Instruct as the weak model and Llama-2-13B-Chat as the strong model. The routing outcome records whether their cached responses disagree. Each prompt is represented by a 138-dimensional context comprising 10 weak-model uncertainty statistics, 64 PCA components of its hidden states, and 64 PCA components of a context-free prompt embedding.

We compare linear-logistic CBPSide (Bartók et al., 2012a; An et al., 2026), SQUARECB.PMSIDE (ours), ETC (Explore-Then-Commit, adapted for apple tasting) (Garivier et al., 2016) with both linear logistic regression and a nonlinear 15-leaf histogram gradient-boosting (HGB) model, Bayesian-logistic PG-TS (Grant & Leslie, 2021), and a random baseline. We consider nine values of ℓ_{01} between 1.8 and 3.3 and evaluate each configuration over 20 paired random permutations. Figure 2a reports the mean realized total cost.

For each value of ℓ_{01} we tune every algorithm separately, selecting the configuration with the lowest total cost. For all methods except the random baseline this amounts to choosing a multiplier $s \in \{0.1, 0.3, 1, 3, 10\}$, which scales the confidence radius in CBPSide, the exploration parameter $s \cdot \kappa$ in both SQUARECB.PMSIDE variants, the initial exploration budget $s \cdot T^{2/3}$ in ETC, and the prior

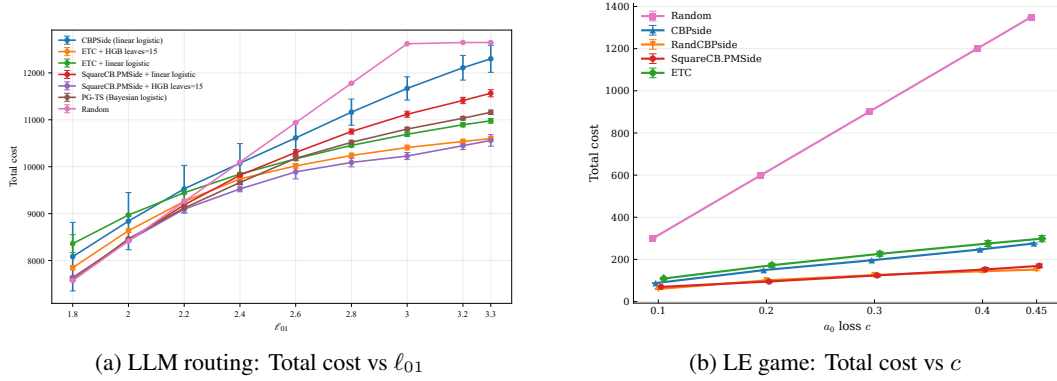


Figure 2: Comparison of algorithms for LLM routing and Label efficient prediction game

standard deviation in PG-TS. The random baseline instead routes each query to the strong model with a fixed probability, independent of context, swept over $\{0, 0.2, 0.4, 0.6, 0.8, 1\}$. All methods, ours included, are therefore tuned with hindsight knowledge of the loss. Further implementation details and results are provided in Appendix E.

Across the evaluated loss settings, both the linear-logistic and nonlinear HGB instantiations of SQUARECB.PMSIDE achieve lower mean realized total cost than CBPSIDE, while the nonlinear HGB variant also outperforms PG-TS. These results illustrate a practical advantage of the oracle-efficient SQUARECB.PMSIDE formulation: it can incorporate flexible nonlinear prediction models without modifying the surrounding routing policy. Furthermore, generally ETC is expected to perform worse due to $O(T^{2/3})$ regret, however here due to limited test data size 12,648 compared to larger dimension 138, it performs competitively. See Figure 2a.

Contextual label-efficient prediction. We conduct an analogous study for the contextual label-efficient prediction game. Local observability requires the expert-query cost to satisfy $c < 0.5$; accordingly, we evaluate five values of c in the range $0.1 \leq c < 0.5$. Because PG-TS is designed specifically for apple-tasting games and is thus inapplicable to this problem, we replace it with RandCBPSide (Heuillet et al., 2024). We generate the data synthetically from a linear logistic model with 10-dimensional context vectors. Further, details are provided in Appendix E. Our method consistently outperforms ETC and CBPSide across the evaluated settings and remains competitive with RandCBPSide, outperforming it for some of the expert-query costs. Moreover, RandCBPSide requires comparatively expensive confidence-width calculations, whereas our method is computationally cheaper. See Figure 2b.

5 CONCLUSION AND FUTURE WORK

Our work gives the first efficient algorithms for stochastic contextual partial monitoring with optimal regret with general function approximation. Experiments on LLM routing and a modified label-efficient game support these guarantees. Our analysis is restricted to non-degenerate games. Lemma G.3, which underlies the $O(1/\gamma)$ bound for locally observable games, relies on non-degeneracy. We conjecture that our SQUARECB.PMSIDE might work for degenerate games as well with minor modification. However, their analysis need a careful approach which is left to future work. Another interesting question is, can our algorithm be generalized to the setting when the loss matrix and the feedback matrix is context-dependent? This seems practically relevant for e.g. LLM routing applications, since the cost of wrongly following the weak model may depend on the importance of the prompt. Furthermore, extending our work to contextual linear partial monitoring (Kirschner et al., 2023) and contextual partial monitoring with infinite outcomes (Lattimore, 2022) are also interesting future directions.

AI USE STATEMENT

In this work, we used generative AI tools for coding for experiments, writing those experiment details and polishing the writing. We did reviewed those coding implementation for correctness. We have not used generative AI tools for writing proof or related works, and the rest of the required disclosure tasks are not applicable to this work. We have reviewed all AI-assisted work. We take responsibility

for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work develops learning algorithms for routing queries among large language models. Such methods may reduce inference costs and computational resource consumption while improving access to capable models. However, routing decisions may propagate or amplify biases present in the underlying models and data, potentially producing unequal performance across demographic groups, languages, or application domains. Furthermore, queries may contain sensitive information, particularly when routed to externally hosted models. Our methods optimize a specified performance objective and do not independently guarantee fairness, privacy, reliability, or safety. Practical deployments should therefore incorporate appropriate data protection, subgroup evaluation, human oversight, and fallback mechanisms.

REPRODUCIBILITY STATEMENT

We provide complete definitions, assumptions, algorithm descriptions, and regret analyses in the main paper and appendices. The experimental section reports the datasets, preprocessing procedures, model configurations, baselines, evaluation metrics, and hyperparameters used in our study. We will release the implementation, experiment configurations, and evaluation scripts later to facilitate reproduction of our results, subject to the licenses and access conditions of the underlying datasets and language models. We also report results across multiple random seeds to account for experimental variability.

ACKNOWLEDGMENTS

This work was partly supported by the National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) in 2026 (Grant No. RS-2026-25612350, Project Title: Development of an End-to-End Design and Validation Framework for Mathematically Principled Next-Generation AI Architectures and Learning Algorithms, contribution 45%). This work was partly supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (RS-2019-II191906, Artificial Intelligence Graduate School Program(POSTECH), contribution 5%; RS-2026-25616313, AI Star Fellowship(POSTECH, Interdisciplinary Program), contribution 50%). This work was supported in part by the National Science Foundation under grant CCF-2327013, IIS-2440266 (CAREER) and Meta Platforms, Inc.

REFERENCES

- Alekh Agarwal, Sham Kakade, Akshay Krishnamurthy, and Wen Sun. Flambe: Structural complexity and representation learning of low rank mdps. *Advances in neural information processing systems*, 33:20095–20107, 2020.
- Byeongwoo An, Kapilan Balagopalan, Hyowon Wi, Sehwa Jeong, Jiun Jeong, Kyoungseok Jang, Noseong Park, Gi-Soo Kim, and Kwang-Sung Jun. Efficient algorithms for contextual apple tasting with log-loss. In *ICML 2026 DEMO workshop*, 2026.
- Peter Auer, Nicolò Cesa-Bianchi, Yoav Freund, and Robert E Schapire. The Nonstochastic Multiarmed Bandit Problem. *SIAM J. Comput.*, 32(1):48–77, jan 2003. ISSN 0097-5397. doi: 10.1137/S0097539701398375.
- Gábor Bartók and Csaba Szepesvári. Partial monitoring with side information. In *International Conference on Algorithmic Learning Theory*, pp. 305–319. Springer, 2012.
- Gábor Bartók et al. Partial monitoring with side information. In *In Proc. ALT*, 2012a.
- Gábor Bartók et al. An adaptive algorithm for finite stochastic partial monitoring. In *Proc. ICML*, 2012b.
- Gábor Bartók, Dávid Pál, and Csaba Szepesvári. Minimax regret of finite partial-monitoring games in stochastic environments. In Sham M. Kakade and Ulrike von Luxburg (eds.), *Proceedings of the 24th Annual Conference on Learning Theory*, volume 19 of *Proceedings of Machine Learning Research*, pp. 133–154, Budapest, Hungary, 09–11 Jun 2011. PMLR. URL <https://proceedings.mlr.press/v19/bartok11a.html>.

- Lingjiao Chen, Matei Zaharia, and James Zou. Frugalgpt: How to use large language models while reducing cost and improving performance. *arXiv preprint arXiv:2305.05176*, 2023.
- Dylan J Foster and Alexander Rakhlin. Beyond UCB: Optimal and efficient contextual bandits with regression oracles. *Proceedings of the International Conference on Machine Learning (ICML)*, 2020.
- Dylan J Foster, Sham M Kakade, Jian Qian, and Alexander Rakhlin. The Statistical Complexity of Interactive Decision Making. *CoRR*, abs/2112.1, 2021.
- Dylan J Foster, Alexander Rakhlin, Ayush Sekhari, and Karthik Sridharan. On the complexity of adversarial decision making. *Advances in Neural Information Processing Systems*, 35:35404–35417, 2022.
- Aurélien Garivier, Tor Lattimore, and Emilie Kaufmann. On explore-then-commit strategies. In *Advances in Neural Information Processing Systems*, volume 29, pp. 784–792. Curran Associates, Inc., 2016.
- James A Grant and David S Leslie. Apple tasting revisited: Bayesian approaches to partially monitored online binary classification. *arXiv preprint arXiv:2109.14412*, 2021.
- Elad Hazan, Amit Agarwal, and Satyen Kale. Logarithmic Regret Algorithms for Online Convex Optimization. *Mach. Learn.*, 69(2-3):169–192, 2007.
- David Helmbold et al. Some label efficient learning results. In *In Proc. CoLT*, 1997.
- David P Helmbold, Nicholas Littlestone, and Philip M Long. Apple tasting. *Information and Computation*, 161(2):85–139, 2000.
- Maxime Heuillet, Ola Ahmad, and Audrey Durand. Randomized confidence bounds for stochastic partial monitoring, 2024.
- Johannes Kirschner, Tor Lattimore, and Andreas Krause. Linear partial monitoring for sequential decision-making: Algorithms, regret bounds and applications. *JMLR*, 2023.
- Robert Kleinberg and Tom Leighton. The value of knowing a demand curve: Bounds on regret for online posted-price auctions. In *44th Annual IEEE Symposium on Foundations of Computer Science, 2003. Proceedings.*, pp. 594–605. IEEE, 2003.
- Junpei Komiyama, Junya Honda, Hisashi Kashima, and Hiroshi Nakagawa. Regret lower bound and optimal algorithm in dueling bandit problem. In *Conference on learning theory*, pp. 1141–1154. PMLR, 2015.
- Tor Lattimore. Minimax regret for partial monitoring: Infinite outcomes and rustichini’s regret. In *Conference on Learning Theory*, pp. 1547–1575. PMLR, 2022.
- Tor Lattimore and Andras Gyorgy. Mirror descent and the information ratio. In *Conference on Learning Theory*, pp. 2965–2992. PMLR, 2021.
- Tor Lattimore and Csaba Szepesvári. An information-theoretic approach to minimax regret in partial monitoring. In *Conference on Learning Theory*, pp. 2111–2139. PMLR, 2019.
- Tor Lattimore and Csaba Szepesvári. Exploration by optimisation in partial monitoring. In *Conference on Learning Theory*, pp. 2488–2515. PMLR, 2020.
- Tor Lattimore and Csaba Szepesvári. *Bandit Algorithms*. Cambridge University Press, 2020.
- David A Levin and Yuval Peres. *Markov Chains and Mixing Times*. American Mathematical Society, Providence, RI, 2 edition, 2017. ISBN 978-1-4704-2962-1.
- Yubo Li, Yidi Miao, Ramayya Krishnan, and Rema Padman. Jev-as-a-judge: Accept when confident, escalate when unsure. *arXiv preprint arXiv:2609.26550*, 2026.

- Isaac Ong, Amjad Almahairi, Vincent Wu, Wei-Lin Chiang, Tianhao Wu, Joseph E Gonzalez, M Kadous, and Ion Stoica. Routellm: Learning to route llms from preference data. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 34433–34448, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/5503a7c69d48a2f86fc00b3dc09de686-Paper-Conference.pdf.
- Hao Qin and Chicheng Zhang. Optimal and efficient contextual combinatorial semi-bandits with general function approximation, 2026a. URL <https://arxiv.org/abs/2607.13686>.
- Hao Qin and Chicheng Zhang. Taming the monster every context: Complexity measure and unified framework for offline-oracle efficient contextual bandits, 2026b. URL <https://arxiv.org/abs/2602.09456>.
- David Sculley. Practical learning from one-sided feedback. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 609–618, 2007.
- David Simchi-Levi and Yunzong Xu. Bypassing the monster: A faster and simpler optimal algorithm for contextual bandits under realizability. *Mathematics of Operations Research*, 47(3):1904–1931, 2022.
- Adish Singla, Ian Lienert, Gábor Bartók, and Andreas Krause. Contextual procurement in online crowdsourcing markets. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 2, pp. 58–59, 2014.
- A. Tsybakov. Introduction to nonparametric estimation. In *Springer Series in Statistics*, 2008. URL <https://api.semanticscholar.org/CorpusID:42933599>.
- Vladimir G Vovk. A game of prediction with expert advice. In *Proceedings of the eighth annual conference on Computational learning theory*, pp. 51–60, 1995.
- Yunbei Xu and Assaf Zeevi. Upper counterfactual confidence bounds: a new optimism principle for contextual bandits. *arXiv preprint arXiv:2007.07876*, 2020.
- Mengxiao Zhang, Yuheng Zhang, Olga Vrousitou, Haipeng Luo, and Paul Mineiro. Practical contextual bandits with feedback graphs. *Advances in Neural Information Processing Systems*, 36: 30592–30617, 2023.
- Yinglun Zhu and Paul Mineiro. Contextual bandits with smooth regret: Efficient learning in continuous action spaces. In *International Conference on Machine Learning*, pp. 27574–27590. PMLR, 2022.
- Yinglun Zhu, Dylan J Foster, John Langford, and Paul Mineiro. Contextual bandits with large action spaces: Made practical. In *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 27428–27453, 2022.

Appendix

Appendix A	Additional Discussions on Important Related Works	15
A.1	Exploration-by-optimization (ExO) versus our approach	15
A.2	Relation to the decision-estimation coefficient	16
Appendix B	Online and offline estimation oracles	17
B.1	Online Oracles	17
B.2	Offline estimation oracle	19
Appendix C	Analysis of E2D.PMSIDE	21
C.1	Proof of Theorem 3.1	21
C.2	Proofs for Algorithm 2 (ϵ -GREEDY.PMSIDE)	23
C.3	Proofs for Algorithm 3 (SQUARECB.PMSIDE)	25
Appendix D	Contextual partial monitoring: reduction to offline regression	28
D.1	OE2D.PMSIDE	28
D.2	Regret analysis	29
Appendix E	Empirical studies	35
Appendix F	Algorithm for L_q -decreasing tree construction	38
Appendix G	Auxiliary Lemmata	38

A ADDITIONAL DISCUSSIONS ON IMPORTANT RELATED WORKS

A.1 EXPLORATION-BY-OPTIMIZATION (ExO) VERSUS OUR APPROACH.

Lattimore & Szepesvári [2020] propose the exploration-by-optimization (ExO) algorithm for adversarial non-contextual partial monitoring. They first form an exponentially weighted benchmark policy λ_t over the Pareto optimal actions from unbiased loss estimates,

$$\lambda_{ta} = \frac{\mathbb{1}_{\Pi}(a) \exp\left(-\eta \sum_{s=1}^{t-1} \hat{y}_{sa}\right)}{\sum_{b \in \Pi} \exp\left(-\eta \sum_{s=1}^{t-1} \hat{y}_{sb}\right)},$$

where $\mathbb{1}_{\Pi}$ is the indicator of Π and $\hat{y}_t = G_t(A_t, \sigma_t)/p_{tA_t}$ is an importance-weighted loss estimate, with G_t (Equation (10)) obtained by solving Equation (4) at round t . This generalizes the EXP3 algorithm [Auer et al., 2003] for adversarial bandits.

They then improve this benchmark by solving

$$\underset{G \in \mathcal{H}, p \in \mathcal{P}}{\text{minimise}} \quad \max_{o \in [d]} \left[(p - \lambda)^\top \mathbf{L}e_o + \text{bias}_\lambda(G; o) + \frac{1}{\eta} \sum_{a=1}^k p_a \Psi_\lambda\left(\frac{\eta G(a, \Phi_{ao})}{p_a}\right) \right], \quad (4)$$

where η plays the role of our γ , $\Psi_q(z) = \langle q, \exp(-z) + z - 1 \rangle$, and

$$\text{bias}_\lambda(G; o) = \left\langle \lambda, \mathbf{L}e_o - \sum_{a=1}^k G(a, \Phi_{ao}) \right\rangle + \max_{c \in \Pi} \left(\sum_{a=1}^k G(a, \Phi_{ao})_c - L_{co} \right).$$

We may write this equivalently, following the notation of Lattimore & Gyorgy [2021], as

$$\underset{G \in \mathcal{H}, p \in \mathcal{P}}{\text{minimise}} \quad \max_{o \in [d], c \in \Pi} \left[(p - \lambda)^\top \mathbf{L}e_o + \left\langle e_c - p, \sum_{a=1}^k G(a, \Phi_{ao}) - \mathbf{L}e_o \right\rangle + \frac{1}{\eta} \sum_{a=1}^k p_a \Psi_\lambda\left(\frac{\eta G(a, \Phi_{ao})}{p_a}\right) \right].$$

This step reallocates probability away from the benchmark policy and toward actions whose feedback is informative. Writing $\text{opt}_\lambda(\eta)$ for the optimal value of Equation (4), the analogue of our $\mathcal{V}_\gamma$, Lattimore & Szepesvári [2020] bound it for both locally and globally observable games without exhibiting an explicit or approximate solution to the optimization problem.

They first restrict the class of estimators G . Let

$$\mathcal{H}_0 = \left\{ G \in \mathcal{H} : \left\langle e_b - e_c, \sum_{a=1}^k G(a, \Phi_{ao}) - \mathbf{L}e_o \right\rangle = 0 \text{ for all } b, c \right\}$$

denote the unbiased estimators, for which $\text{bias}_\lambda(G; o) = 0$ for every o . The minimax objective is then bounded above by

$$\min_{G \in \mathcal{H}_0, p \in \mathcal{P}} \max_{o \in [d]} \left[(p - \lambda)^\top \mathbf{L}e_o + \frac{1}{\eta} \sum_{a=1}^k p_a \Psi_\lambda\left(\frac{\eta G(a, \Phi_{ao})}{p_a}\right) \right],$$

and, since the bracketed expression is linear in e_o , maximizing over outcomes is equivalent to maximizing over outcome distributions:

$$\min_{G \in \mathcal{H}_0, p \in \mathcal{P}} \max_{q \in \mathcal{D}} \left[(p - \lambda)^\top \mathbf{L}q + \frac{1}{\eta} \sum_{o=1}^d q_o \sum_{a=1}^k p_a \Psi_\lambda\left(\frac{\eta G(a, \Phi_{ao})}{p_a}\right) \right]. \quad (5)$$

Equation (5) is concave in q and convex in (G, p) , so to (nonconstructively) upper bound the value of Equation (5), Sion's minimax theorem permits exchanging the two operations. Writing $F(G, p, q)$ for the bracketed expression,

$$\min_{G \in \mathcal{H}_0, p \in \mathcal{P}} \max_{q \in \mathcal{D}} F(G, p, q) = \max_{q \in \mathcal{D}} \min_{G \in \mathcal{H}_0, p \in \mathcal{P}} F(G, p, q).$$

To upper bound this expression, it now suffices to choose G and p as functions of q and λ : they build an Lq -decreasing tree, take G to be the loss estimator induced by that tree, and set $p = W_q(\lambda)$, where W_q is the water-transfer operator (see Lemma G.5 for a recap). The operator redistributes the benchmark policy λ so that the resulting p incurs no greater loss under q , while allocating probability that increases toward the root of the tree.

In contrast, *given an outcome distribution q* , our exploration strategy approximately solves the following optimization problem:

$$\min_{p \in \mathcal{P}} \max_{\lambda \in \Delta(\Pi)} (p - \lambda)^\top \mathbf{L} \hat{q} + \frac{1}{\gamma} \cdot \text{Coverage}(p, \lambda)$$

Given that for any $G \in \mathcal{H}_0$, Lemma C.1 implies that $\text{Coverage}(p, \lambda) \leq \sum_a \sum_b \frac{\lambda_b + p_b}{p_a} \|G_b(a, \cdot)\|_\infty^2$, and thus it suffices to find exploration distribution p that solves,

$$\min_{p \in \mathcal{P}} \max_{\lambda \in \Delta(\Pi)} (p - \lambda)^\top \mathbf{L} \hat{q} + \frac{1}{\gamma} \cdot \left(\sum_a \sum_b \frac{\lambda_b + p_b}{p_a} \|G_b(a, \cdot)\|_\infty^2 \right) \quad (6)$$

Comparing Equation (5) with Equation (6), the two objectives share a shape: each consists of a greediness term of the form $(p - \lambda)^\top \mathbf{L} \hat{q}$ and a coverage term measuring how well p covers λ . They differ in three respects:

- First, in the minimax players: we minimize over exploration policies against a maximizing comparator policy λ , whereas Equation (5) minimizes against a maximizing adversarial outcome distribution q .
- Second, in the form of the coverage term. The one in Equation (5) is exponential, $\Psi_\lambda(\eta G(a, \Phi_{ao})/p_a)$, and Lattimore & Szepesvári [2020] relax it to $\eta^2 \sum_b \lambda_b \|G(a, \cdot)_b\|_\infty^2 / p_a^2$, at the cost of requiring p to be bounded below entrywise by $\eta \max_{a,b} \|G(a, \cdot)_b\|_\infty$.
- Third, in what the analysis delivers. ExO is tractable but gives no explicit exploration policy p . For our objective we show that the inverse-gap-weighting rule of SQUARECB.PMSIDE certifies a small value of $\mathcal{V}_\gamma$ directly: it already allocates probability increasing toward the root of an $\mathbf{L} \hat{q}$ -decreasing tree, so neither the optimization over G nor the water-transfer step is needed, and the in-tree only appears in the analysis.

We also refer the reader to Table 1 for a summary.

Table 1: Comparison with exploration-by-optimization [Lattimore & Szepesvári, 2020]. $\uparrow$ indicates a better result. NA indicates not applicable

	ExO	Ours
Side information	No	Yes $\uparrow$
Outcome distribution	Adversarial $\uparrow$	Stochastic
Objective	$\text{opt}_q(\eta)$	$\mathcal{V}_\gamma$
Minimax players	Policy vs. adversarial outcome x	Policy vs. comparator policy λ
Optimization variables	G and p jointly	p only $\uparrow$
Per-round computation	Exact solution of Equation (4)	IGW for locally observable; ϵ -greedy for globally observable $\uparrow$
In-tree	Optimized at every round	Analysis only $\uparrow$
Water-transfer	Required	Not required $\uparrow$
Oracle-efficient	No	Efficient $\uparrow$
Offline oracle compatibility	NA	Compatible $\uparrow$
Locally observable regret	—	$\sqrt{k}$ smaller $\uparrow$

A.2 RELATION TO THE DECISION-ESTIMATION COEFFICIENT

The decision-estimation coefficient (DEC) was introduced by Foster et al. [2021] as a complexity measure for stochastic sequential decision making. Foster et al. [2022] extend it to the adversarial setting, showing that the DEC applied to the convex hull of the model class characterizes the fundamental statistical complexity of that setting as well. On the algorithmic side they propose ExO⁺, a high-probability variant of the exploration-by-optimization algorithm of Lattimore & Szepesvári [2020]; Lattimore & Gyorgy [2021]. Whereas the original ExO gives in-expectation guarantees, ExO⁺ gives high-probability guarantees that hold against adaptive adversaries. Like ExO, the exploration rule and the loss estimator are defined through an optimization problem rather than a closed-form policy.

Their analysis does not bound the ExO^+ or DEC objective directly for locally and globally observable games. Instead they establish an equivalence, for general divergences D , between the convexified DEC, a parameterized variant of the generalized information ratio of Lattimore & Gyorgy [2021], and the ExO^+ objective. Since Lattimore & Szepesvári [2019]; Lattimore & Gyorgy [2021]; Lattimore [2022] bound the regret for locally and globally observable games in terms of the generalized information ratio under the KL divergence, in agreement with the classification theorem [Bartók et al., 2011; Lattimore & Szepesvári, 2020], those bounds transfer to ExO^+ through this equivalence. The equivalence itself holds for general D , but the information-ratio bounds are established for the KL divergence; we are not aware of corresponding bounds for other divergences, including the squared Hellinger.

To summarize, Foster et al. [2022]’s result implicitly implies that the E2D algorithm Foster et al. [2021] has optimal order-in- T regret for stochastic contextual partial monitoring. However, as we discussed in Section 2.3, the E2D algorithm may be computationally inefficient. Our paper addresses the computational challenge. For locally observable games we give a closed-form inverse-gap-weighting rule (SQUARECB.PMSIDE), and for globally observable games a closed-form ε -greedy rule (ε -GREEDY.PMSIDE). Both extend to be compatible with offline estimation oracles, because the exploration term of our objective, the coverage, does not involve the current estimate.

B ONLINE AND OFFLINE ESTIMATION ORACLES

B.1 ONLINE ORACLES

In this section, we provide examples for the online estimation oracles Alg_{On} and prove their bounds for both finite and infinite model classes.

Lemma B.1 (Finite-class online Hellinger guarantee via Vovk’s algorithm). Suppose that the candidate conditional outcome-model class

$$\mathcal{Q} \subseteq \{q : \mathcal{X} \rightarrow \Delta(d)\}$$

is finite and realizable, so that $q^* \in \mathcal{Q}$ and the oracles are allowed to output predictors in $\text{co}(\mathcal{Q})$. Then there exists a contextual online estimation oracle Alg_{On} such that,

$$\overline{\text{Reg}}_{\text{On}}^{\text{H}}(\mathcal{Q}, T, \delta) = 2 \log \left(\frac{|\mathcal{Q}|}{\delta} \right).$$

The contextual online estimation oracle can be instantiated by applying Vovk’s [Vovk, 1995] aggregating algorithm to the conditional outcome models in $\mathcal{Q}$.

Proof. Define $\mathcal{G} \subseteq \Delta(m)$ as a set of functions $g_q : \mathcal{X} \times [k] \rightarrow \Delta(m)$ induced by q , formally

$$\mathcal{G} = \{g_q : g_q(\sigma \mid x, a) = \mathbf{S}_a q(x)[\sigma] \quad \forall q \in \mathcal{Q}\}$$

Hence,

$$|\mathcal{G}| \leq |\mathcal{Q}|$$

Specifically,

$$\begin{aligned} g_{q^*}(\sigma \mid x_t, A_t) &= \mathbb{P}(\sigma_t = \sigma \mid x_t, A_t) \\ &= S_{A_t} q^*(x_t)[\sigma] \end{aligned}$$

Vovk’s algorithm calculate the $g_{\hat{q}_t}$ in the following way.

Let,

$$\begin{aligned} \ell_{\log}^{(t)}(g_q) &:= \log \left(\frac{1}{g_q(\sigma_t \mid x_t, A_t)} \right) \\ &= \log \left(\frac{1}{(S_{A_t} q(x_t))[\sigma_t]} \right). \end{aligned}$$

Then $\forall g_q \in \mathcal{G}$, the algorithm calculates a probability simplex $b_t \in \Delta(|\mathcal{G}|)$ as follows.

$$b_t(g_q) \propto \exp\left(-\sum_{s=1}^{t-1} \ell_{\log}^{(s)}(g_q)\right)$$

Estimate $g_{\hat{q}_t}$

$$g_{\hat{q}_t}(\cdot) = \mathbb{E}_{g \sim b_t} [g(\cdot)]$$

Let the corresponding $\hat{q}_t$ such that,

$$\hat{q}_t(\cdot) = \mathbb{E}_{g \sim b_t} [q(\cdot)].$$

Then Vovk's algorithm guarantees,

$$\sum_{t=1}^T \log \frac{1}{g_{\hat{q}_t}(\sigma_t | x_t, a_t)} - \sum_{t=1}^T \log \frac{1}{g_{q^*}(\sigma_t | x_t, a_t)} \leq \log |\mathcal{G}|$$

Then, by applying Lemma A.14 from Foster et al. [2021], the output estimator of Vovk's aggregating algorithm $g_{\hat{q}_t}$ satisfies, with probability of at least $1 - \delta$,

$$\sum_{t=1}^T \mathbb{E}_{t-1} \left[D_{\text{H}}^2(g_{\hat{q}_t}(\cdot | x_t, A_t) | g_{q^*}(\cdot | x_t, A_t)) \right] \leq 2 \log \frac{|\mathcal{G}|}{\delta} \quad (7)$$

transforming this to our case, with probability at least $1 - \delta$, $\hat{q}_t$ satisfies,

$$\sum_{t=1}^T \mathbb{E}_{a \sim p_t} \left[D_{\text{H}}^2(\mathcal{S}_a \hat{q}_t(x_t) | \mathcal{S}_a q^*(x_t)) \right] \leq 2 \log \frac{|\mathcal{Q}|}{\delta}.$$

Remark 1. Define filtration $\mathcal{H}$ as,

$$\mathcal{H}_{t-1} = \sigma(x_1, A_1, \sigma_1, \dots, x_{t-1}, A_{t-1}, \sigma_{t-1})$$

Equation 7 also implies that, if $x_1, \dots, x_T \stackrel{\text{i.i.d.}}{\sim} \mathcal{D}_x$ and $p_t, \hat{q}_t$ are $\mathcal{H}_{t-1}$ measurable, then with probability of at least $1 - \delta$,

$$\sum_{t=1}^T \mathbb{E}_{x \sim \mathcal{D}_x, a \sim p_t(\cdot | x)} \left[D_{\text{H}}^2(\mathcal{S}_a \hat{q}_t(x) | \mathcal{S}_a q^*(x)) \right] \leq 2 \log \frac{|\mathcal{Q}|}{\delta}.$$

□

Lemma B.2 (Infinite-class online Hellinger guarantee via contextual smoothed aggregation). Suppose that the candidate conditional outcome-model class is,

$$\mathcal{Q} \subseteq \{q : \mathcal{X} \rightarrow \Delta(d)\},$$

is possibly infinite but has finite covering numbers under the uniform Hellinger metric and that the model is realizable $q^* \in \mathcal{Q}$ and the oracles are allowed to output predictors in $\text{co}(\mathcal{Q})$.

Define $\mathcal{G} \subseteq \Delta(m)$ as a set of functions $g_q : \mathcal{X} \times [k] \rightarrow \Delta(m)$ induced by q , formally

$$\mathcal{G} = \{g_q : g_q(\sigma | x, a) = \mathcal{S}_a q(x)[\sigma] \mid q \in \mathcal{Q}\}$$

Moreover, let us define an ϵ -cover $\mathcal{J}$, such that,

$$\forall g_q \in \mathcal{G} \quad \exists g'_q \in \mathcal{J} \text{ such that } \sup_{x \in \mathcal{X}, a \in [k]} D_{\text{H}}^2(g'_q(\cdot | x, a) | g_q(\cdot | x, a)) \leq \epsilon^2$$

Furthermore, let $\mathcal{N}(\mathcal{G}, \epsilon)$ denote the size of the smallest such cover. Define,

$$\text{est}(\mathcal{G}, T) := \inf_{\epsilon \geq 0} \{\log(\mathcal{N}(\mathcal{G}, \epsilon)) + \epsilon^2 T\}$$

Then there exists a contextual online estimation oracle Alg_{On} such that,

$$\overline{\text{Reg}}_{\text{On}}^{\text{H}}(\mathcal{Q}, T, \delta) \leq O\left(\log(2m^2T) \cdot \text{est}(\mathcal{G}, T) + \log^2(2m^2T) \log \frac{1}{\delta}\right).$$

The contextual online estimation oracle can be instantiated by applying an online aggregation algorithm Vovk [1995]; Foster et al. [2021] to a suitably discretized and smoothed approximation of the parameter space $\mathcal{Q}$.

Proof. The regularity assumptions required in Assumption A.2 of Foster et al. [2021] are already satisfied by setting $B = \max\{m, e\}$

1. Number of possible outcome for a context is upper bounded by $m \leq B$
2. $g_q(\sigma \mid x, a)$ is a probability simplex ($1 \leq B$).

Smoothed Vovk's algorithm calculate the $g_{\hat{q}_t}$ in the following way, let,

$$h(\sigma \mid x, a) = (1 - \alpha)g_q(\sigma \mid x, a) + \frac{\alpha}{m}$$

then $\forall g_q \in \mathcal{J}$ calculate the a probability simplex $b_t \in \Delta(|\mathcal{J}|)$ as follows

$$b_t(g_q) \propto \exp\left(-\sum_{s=1}^{t-1} \ell_{\log}^{(s)}(h)\right)$$

Estimate $\hat{g}_t$

$$g_{\hat{q}_t} = \mathbb{E}_{g_q \sim b_t} [g_q]$$

Hence,

$$\hat{q}_t = \mathbb{E}_{q \sim b_t} [q]$$

Then by applying Lemma A.16 from Foster et al. [2021], with probability of at least $1 - \delta$,

$$\sum_{t=1}^T \mathbb{E}_{t-1} \left[D_{\text{H}}^2(g_{\hat{q}_t}(\cdot \mid x_t, A_t) \mid g_{q^*}(\cdot \mid x_t, A_t)) \right] \leq O\left(\log(2m^2T) \cdot \text{est}(\mathcal{G}, T) + \log^2(2m^2T) \log \frac{1}{\delta}\right) \quad (8)$$

Let $\hat{q}_t \in \mathcal{Q}$ be the corresponding solution, with probability of at least $1 - \delta$, $\hat{q}_t$ satisfies,

$$\sum_{t=1}^T \mathbb{E}_{a \sim p_t} \left[D_{\text{H}}^2(\mathcal{S}_a \hat{q}_t(x_t) \mid \mathcal{S}_a q^*(x_t)) \right] \leq O\left(\log(2m^2T) \cdot \text{est}(\mathcal{G}, T) + \log^2(2m^2T) \log \frac{1}{\delta}\right).$$

Remark 2. Define filtration $\mathcal{H}$ as,

$$\mathcal{H}_{t-1} = \sigma(x_1, A_1, \sigma_1, \dots, x_{t-1}, A_{t-1}, \sigma_{t-1})$$

Equation 7 also implies that, if $x_1, \dots, x_T \stackrel{\text{i.i.d.}}{\sim} \mathcal{D}_x$ and $p_t, \hat{q}_t$ are $\mathcal{H}_{t-1}$ measurable, then with probability of at least $1 - \delta$,

$$\sum_{t=1}^T \mathbb{E}_{x \sim \mathcal{D}_x, a \sim p_t(\cdot \mid x)} \left[D_{\text{H}}^2(\mathcal{S}_a \hat{q}_t(x) \mid \mathcal{S}_a q^*(x)) \right] \leq O\left(\log(2m^2T) \cdot \text{est}(\mathcal{G}, T) + \log^2(2m^2T) \log \frac{1}{\delta}\right).$$

□

B.2 OFFLINE ESTIMATION ORACLE

In this section, first we define *offline estimation oracle* and we provide examples for the offline estimation oracles Alg_{Off} and prove their estimation bounds.

Definition B.1 (Offline estimation oracle). An *offline estimation oracle* Alg_{Off} is an algorithm that estimates the unknown parameter $q^* \in \mathcal{Q}$ from the observed data in a batch fashion. That is, after observing t rounds, given the feedback symbols collected so far, the oracle produces an estimate

$$\hat{q} = \text{Alg}_{\text{Off}}\left(\left\{(A_s, x_s, \sigma_s)\right\}_{s=1}^t\right).$$

Assumption B.1 (Hellinger-closeness of the offline estimation oracle). For any $t \in \mathbb{N}$, let $x_1, \dots, x_t \stackrel{\text{i.i.d.}}{\sim} \mathcal{D}_x$ and let $\hat{q}$ denote the estimate returned by Alg_{Off} on the resulting sample. Define $\text{Reg}_{\text{Off}}^H(\mathcal{Q}, t)$, such that

$$\mathbb{E}_{x \sim \mathcal{D}_x, a \sim p(\cdot|x)} \left[D_H^2(\mathcal{S}_a \hat{q}(x) \parallel \mathcal{S}_a q^*(x)) \right] = \frac{\text{Reg}_{\text{Off}}^H(\mathcal{Q}, t)}{t}$$

where p denotes the behavioral policy used to collect the sample, and the expectation is taken over a fresh draw (x, a) , conditionally on $\hat{q}$. We assume that the offline estimation oracle Alg_{Off} guarantees, for any $\delta \in (0, 1)$, that with probability at least $1 - \delta$,

$$\text{Reg}_{\text{Off}}^H(\mathcal{Q}, t) \leq \overline{\text{Reg}}_{\text{Off}}^H(\mathcal{Q}, t, \delta),$$

where $\overline{\text{Reg}}_{\text{Off}}^H(\mathcal{Q}, t, \delta)$ is a known, deterministic upper bound.

Having established in Appendix B.1 the existence of online oracles with favorable guarantees for both finite and infinite model classes, we can always obtain offline oracles for both by online-to-batch conversion. Beyond this generic construction, we also show that the conditional maximum-likelihood estimator satisfies Assumption B.1 with a favorable bound $\overline{\text{Reg}}_{\text{Off}}^H(\mathcal{Q}, t, \delta)$.

Lemma B.3 (Existence of a contextual offline estimation oracle for a finite class). Suppose that the candidate conditional outcome-model class

$$\mathcal{Q} \subseteq \{q : \mathcal{X} \rightarrow \Delta(d)\}$$

is finite and realizable, so that $q^* \in \mathcal{Q}$. Then there exists a contextual offline estimation oracle Alg_{Off} such that Assumption B.1 is satisfied with

$$\overline{\text{Reg}}_{\text{Off}}^H(\mathcal{Q}, t, \delta) = 2 \log \left(\frac{|\mathcal{Q}|}{\delta} \right).$$

The contextual offline estimation oracle can be instantiated using the conditional maximum-likelihood estimator

$$\hat{q}_t \in \arg \max_{q \in \mathcal{Q}} \sum_{s=1}^t \log \left((S_{A_s} q(x_s))[\sigma_s] \right).$$

Proof. Let

$$\begin{aligned} z_s &= (x_s, A_s) & (\text{Where } x_s \sim \mathcal{X}(\text{i.i.d}) \text{ and } A_s \sim p_s(z_{1:s-1}, y_{1:s-1}, x_s)) \\ y_s &= \sigma_s \end{aligned}$$

Hence,

$$z_s \sim \mathcal{D}_s(z_{1:s-1}, y_{1:s-1}) \quad (\text{for some distribution } \mathcal{D}_s)$$

Define

$$\begin{aligned} f^*(z_s, y_s) &= \mathbb{P}(y_s \mid z_s) \\ &= \mathbb{P}(\sigma_s \mid A_s, x_s) \\ &= \mathbb{P}(\sigma_s \mid S_{A_s} q^*(x_s)) \\ &= S_{A_s} q^*(x_s)[\sigma] \end{aligned}$$

Furthermore, define

$$\mathcal{F} = \{f : f(z, y) = S_a q(x)[\sigma] \quad \forall q \in \mathcal{Q}\}$$

Hence,

$$|\mathcal{F}| \leq |\mathcal{Q}|$$

Following the work of Agarwal et al. [2020] (section E)

$$\begin{aligned} \hat{f} &:= \arg \max_{f \in \mathcal{F}} \sum_{s=1}^t \log f(z_s, y_s) \\ &= \arg \max_q \sum_{s=1}^t \mathbb{P}(\sigma_s \mid S_{A_s} q(x_s)) \end{aligned}$$

Therefore, let

$$\begin{aligned}\hat{q}_t &= \arg \max_q \sum_{s=1}^t \log \left(\mathbb{P}(\sigma_s \mid x_s, A_s, q) \right) \\ &= \arg \max_q \sum_{s=1}^t \log (S_{A_s} \cdot q(x_s)[\sigma_s])\end{aligned}$$

then,

$$\hat{f}(z, y) = \mathbf{S}_a \hat{q}_t(x)$$

By Theorem 21 of Agarwal et al. [2020], this maximum likelihood solution, satisfies the following,

$$\sum_{s=1}^t \mathbb{E}_{z \sim D_s} \|\hat{f}(z, \cdot) - f^*(z, \cdot)\|_{\text{TV}}^2 \leq 2 \log \left(\frac{|\mathcal{F}|}{\delta} \right)$$

Although Theorem 21 is stated in terms of squared total variation distance, Remark 23 observes that its proof directly establishes the corresponding squared-Hellinger bound.

$$\sum_{s=1}^t \mathbb{E}_{z \sim D_s} \left[D_{\text{H}}^2 \left(\hat{f}(z, \cdot) \mid f^*(z, \cdot) \right) \right] \leq 2 \log \left(\frac{|\mathcal{F}|}{\delta} \right)$$

Therefore, in our case

$$\sum_{s=1}^t \mathbb{E}_{x \sim \mathcal{X}, a \sim p_s} \left[D_{\text{H}}^2 \left(\mathbf{S}_a \cdot \hat{q}_t(x) \mid \mathbf{S}_a \cdot q^*(x) \right) \right] \leq 2 \log \left(\frac{|\mathcal{Q}|}{\delta} \right)$$

□

C ANALYSIS OF E2D.PMSIDE

C.1 PROOF OF THEOREM 3.1

Theorem 3.1 (Regret of E2D.PMSIDE). Suppose Alg_{On} has a Hellinger regret bound $\overline{\text{Reg}}_{\text{On}}^{\text{H}}(\mathcal{Q}, T, \delta)$. Fix $\gamma > 0$. Then, for any $\delta \in (0, 1)$, with probability at least $1 - \delta$,

$$\text{Reg}_T \leq T \cdot \max_{t=1}^T \mathcal{V}_\gamma(\hat{q}_t(x_t), p_t) + \gamma \cdot \overline{\text{Reg}}_{\text{On}}^{\text{H}}(\mathcal{Q}, T, \delta).$$

Proof. Let

$$\lambda_{x_t}^* = \arg \min_{p' \in \Delta(\Pi)} (p')^\top \mathbf{L} q^*(x_t) \quad (\text{The best policy for context } x_t)$$

$$\begin{aligned}\text{Reg}_T &= \sum_{t=1}^T p_t^\top \mathbf{L} q^*(x_t) - (\lambda_{x_t}^*)^\top \mathbf{L} q^*(x_t) \\ &= \sum_{t=1}^T p_t^\top \mathbf{L} \hat{q}_t(x_t) - (\lambda_{x_t}^*)^\top \mathbf{L} \hat{q}_t(x_t) + (\lambda_{x_t}^* - p_t)^\top \mathbf{L} (\hat{q}_t(x_t) - q^*(x_t))\end{aligned}$$

$$\begin{aligned}((\lambda_{x_t}^* - p_t)^\top \mathbf{L} (\hat{q}_t(x_t) - q^*(x_t)))^2 &= \frac{\left((\lambda_{x_t}^* - p_t)^\top \mathbf{L} (\hat{q}_t(x_t) - q^*(x_t)) \right)^2}{\mathbb{E}_{a \sim p_t} \left[D_{\text{H}}^2(\mathbf{S}_a \hat{q}_t(x_t) \mid \mathbf{S}_a q^*(x_t)) \right]} \cdot \left(\mathbb{E}_{a \sim p_t} \left[D_{\text{H}}^2(\mathbf{S}_a \hat{q}_t(x_t) \mid \mathbf{S}_a q^*(x_t)) \right] \right) \\ &\leq \max_{q_1, q_2 \in \Delta(d)} \frac{\left((\lambda_{x_t}^* - p_t)^\top \mathbf{L} (q_1 - q_2) \right)^2}{\mathbb{E}_{a \sim p_t} \left[D_{\text{H}}^2(\mathbf{S}_a q_1 \mid \mathbf{S}_a q_2) \right]} \cdot \left(\mathbb{E}_{a \sim p_t} \left[D_{\text{H}}^2(\mathbf{S}_a \hat{q}_t(x_t) \mid \mathbf{S}_a q^*(x_t)) \right] \right).\end{aligned}$$

Therefore,

$$\begin{aligned}
((\lambda_{x_t}^* - p_t)^\top \mathbf{L}(\hat{q}_t(x_t) - q^*(x_t)))^2 &\leq \text{Coverage}(p_t, \lambda_{x_t}^*) \cdot \left(\mathbb{E}_{a \sim p_t} \left[D_{\text{H}}^2(\mathbf{S}_a \hat{q}_t(x_t) \mid \mathbf{S}_a q^*(x_t)) \right] \right) \\
((\lambda_{x_t}^* - p_t)^\top \mathbf{L}(\hat{q}_t(x_t) - q^*(x_t))) &\leq \sqrt{\text{Coverage}(p_t, \lambda_{x_t}^*) \cdot \left(\mathbb{E}_{a \sim p_t} \left[D_{\text{H}}^2(\mathbf{S}_a \hat{q}_t(x_t) \mid \mathbf{S}_a q^*(x_t)) \right] \right)} \\
&\leq \frac{1}{\gamma} \text{Coverage}(p_t, \lambda_{x_t}^*) + \gamma \left(\mathbb{E}_{a \sim p_t} \left[D_{\text{H}}^2(\mathbf{S}_a \hat{q}_t(x_t) \mid \mathbf{S}_a q^*(x_t)) \right] \right). \quad (\text{AM-GM})
\end{aligned}$$

$$\begin{aligned}
\text{Reg}_T &\leq \sum_{t=1}^T \sup_{\lambda \in \Delta(\Pi)} \left[p_t^\top \mathbf{L} \hat{q}_t(x_t) - \lambda^\top \mathbf{L} \hat{q}_t(x_t) + \frac{1}{\gamma} \text{Coverage}(p_t, \lambda) \right] + \sum_{t=1}^T \gamma \left(\mathbb{E}_{a \sim p_t} \left[D_{\text{H}}^2(\mathbf{S}_a \hat{q}_t(x_t) \mid \mathbf{S}_a q^*(x_t)) \right] \right) \\
&\leq \sum_{t=1}^T \sup_{\lambda \in \Delta(\Pi)} \left[p_t^\top \mathbf{L} \hat{q}_t(x_t) - \lambda^\top \mathbf{L} \hat{q}_t(x_t) + \frac{1}{\gamma} \cdot \text{Coverage}(p_t, \lambda) \right] + \gamma \left(\text{Reg}_{\text{On}}^{\text{H}}(\mathcal{Q}, T) \right).
\end{aligned}$$

Therefore the regret can be bounded as,

$$\begin{aligned}
\text{Reg}_T &\leq \sum_{t=1}^T \mathcal{V}_\gamma(\hat{q}_t(x_t), p_t) + \gamma \left(\text{Reg}_{\text{On}}^{\text{H}}(\mathcal{Q}, T) \right) \\
&\leq T \cdot \max_{t=1}^T \mathcal{V}_\gamma(\hat{q}_t(x_t), p_t) + \gamma \left(\text{Reg}_{\text{On}}^{\text{H}}(\mathcal{Q}, T) \right).
\end{aligned}$$

Hence, with probability of at least $1 - \delta$,

$$\text{Reg}_T \leq T \cdot \max_{t=1}^T \mathcal{V}_\gamma(\hat{q}_t(x_t), p_t) + \gamma \cdot \overline{\text{Reg}}_{\text{On}}^{\text{H}}(\mathcal{Q}, T, \delta).$$

□

Moreover,

Remark 3 (DEC bound). For all $x \in \mathcal{X}, \hat{q} \in \mathcal{Q}, p \in \Delta(d)$,

$$\mathcal{V}_\gamma^{\text{dec}}(\hat{q}(x), \mathcal{Q}_x, p) \leq \mathcal{V}_\gamma(\hat{q}(x), p).$$

Proof.

$$\begin{aligned}
\mathcal{V}_\gamma^{\text{dec}}(\hat{q}(x), \mathcal{Q}_x, p) &= \sup_{q \in \mathcal{Q}_x, \lambda \in \Delta(k)} \left[p^\top \mathbf{L} q(x) - \lambda^\top \mathbf{L} q(x) - \gamma \cdot \mathbb{E}_{a \sim p} \left[D_{\text{H}}^2(\mathbf{S}_a \hat{q}(x) \mid \mathbf{S}_a q(x)) \right] \right] \\
&= \sup_{q \in \mathcal{Q}_x, \lambda \in \Delta(\Pi)} \left[p^\top \mathbf{L} q(x) - \lambda^\top \mathbf{L} q(x) - \gamma \cdot \mathbb{E}_{a \sim p} \left[D_{\text{H}}^2(\mathbf{S}_a \hat{q}(x) \mid \mathbf{S}_a q(x)) \right] \right] \\
&= \sup_{q \in \mathcal{Q}_x, \lambda \in \Delta(\Pi)} \left[p^\top \mathbf{L} \hat{q}(x) - \lambda^\top \mathbf{L} \hat{q}(x) + (\lambda - p)^\top \mathbf{L}(\hat{q}(x) - q(x)) - \gamma \cdot \mathbb{E}_{a \sim p} \left[D_{\text{H}}^2(\mathbf{S}_a \hat{q}(x) \mid \mathbf{S}_a q(x)) \right] \right] \\
&\leq \sup_{q \in \mathcal{Q}_x, \lambda \in \Delta(\Pi)} \left[p^\top \mathbf{L} \hat{q}(x) - \lambda^\top \mathbf{L} \hat{q}(x) + \frac{1}{\gamma} \text{Coverage}(p, \lambda) \right] \\
&= \sup_{\lambda \in \Delta(\Pi)} \left[p^\top \mathbf{L} \hat{q}(x) - \lambda^\top \mathbf{L} \hat{q}(x) + \frac{1}{\gamma} \text{Coverage}(p, \lambda) \right] \\
&= \mathcal{V}_\gamma(\hat{q}(x), p)
\end{aligned}$$

□

C.2 PROOFS FOR ALGORITHM 2 (ϵ -GREEDY.PMSIDE)

Recall that in partial monitoring the loss difference between two actions can be expressed in terms of the feedback from other actions (Definitions 2.2 and 2.3), though the two observability conditions supply that expression differently. The following lemma abstracts both constructions into a single reference witness G and bounds the coverage in terms of it.

Lemma C.1. Suppose there exists a *reference witness function* $G : \mathcal{A} \times \Sigma \rightarrow \mathbb{R}^{\mathcal{A}}$, in the sense that for all $b, c \in \mathcal{A} \subseteq [k]$,

$$\sum_{a \in \mathcal{A}} G_b(a, \Phi_{ao}) - \sum_{a \in \mathcal{A}} G_c(a, \Phi_{ao}) = L_{bo} - L_{co} \quad \forall o \in [d]. \quad (9)$$

where, G_b denotes b -th index of G . Then $\forall p, \lambda \in \Delta(\mathcal{A})$,

$$\text{Coverage}(p, \lambda) \leq 4 \sum_{a \in \mathcal{A}} \frac{(\sum_{b \in \mathcal{A}} (p_b - \lambda_b) \|G_b(a, \cdot)\|_{\infty})^2}{p_a} \leq 8 \sum_{a \in \mathcal{A}} \sum_{b \in \mathcal{A}} \frac{p_b + \lambda_b}{p_a} \|G_b(a, \cdot)\|_{\infty}^2$$

Proof. First, since $D_H^2(u | v) \geq \frac{1}{4} \|u - v\|_1^2$ (Lemma G.2), we have,

$$\text{Coverage}(p, \lambda) = \sup_{q_1, q_2 \in \Delta(d)} \frac{\left((\lambda - p)^{\top} L (q_1 - q_2) \right)^2}{\mathbb{E}_{a \sim p} \left[D_H^2(\mathcal{S}_a q_1 | \mathcal{S}_a q_2) \right]} \leq 4 \sup_{q_1, q_2 \in \Delta(d)} \frac{\left((\lambda - p)^{\top} L (q_1 - q_2) \right)^2}{\sum_{a \in \mathcal{A}} \|\mathcal{S}_a (q_1 - q_2)\|_1^2}$$

Taking a weighted summation of Equation (9) over b, c with weight $p_b \lambda_c$ gives,

$$(p - \lambda)^{\top} \left(\sum_a G(a, \Phi_{ao}) - L_{:,o} \right) = 0, \quad \forall o$$

And thus,

$$\begin{aligned} (p - \lambda)^{\top} L (q_1 - q_2) &= \sum_o (p - \lambda)^{\top} \sum_{a \in \mathcal{A}} G(a, \Phi_{ao}) (q_1 - q_2)_o \\ &= \sum_o (p - \lambda)^{\top} \sum_{a \in \mathcal{A}} \sum_{\sigma} G(a, \sigma) (\mathcal{S}_a)_{\sigma, o} (q_1 - q_2)_o \\ &= (p - \lambda)^{\top} \sum_{\sigma} \sum_{a \in \mathcal{A}} G(a, \sigma) (\mathcal{S}_a (q_1 - q_2))_{\sigma} \\ &\leq \sum_{b \in \mathcal{A}} \sum_{\sigma} \sum_{a \in \mathcal{A}} (p_b - \lambda_b) G_b(a, \sigma) (\mathcal{S}_a (q_1 - q_2))_{\sigma} \end{aligned}$$

And therefore, exchanging the summation by moving $\sum_{\sigma}$ to the innermost, and by triangle inequality, $|\sum_{\sigma} G_b(a, \sigma) \mathcal{S}_a (q_1 - q_2)_{\sigma}| \leq \|G_b(a, \cdot)\|_{\infty} \|\mathcal{S}_a (q_1 - q_2)\|_1$; this implies that

$$\left| (p - \lambda)^{\top} L (q_1 - q_2) \right| \leq \sum_{a \in \mathcal{A}} \sum_{b \in \mathcal{A}} |p_b - \lambda_b| \|G_b(a, \cdot)\|_{\infty} \|\mathcal{S}_a (q_1 - q_2)\|_1.$$

The first inequality of the lemma now follows from the definition of coverage and Cauchy-Schwarz inequality.

The second inequality follows from that

$$\begin{aligned} \left(\sum_{b \in \mathcal{A}} (p_b - \lambda_b) \|G_b(a, \cdot)\|_{\infty} \right)^2 &\leq 2 \left(\sum_{b \in \mathcal{A}} p_b \|G_b(a, \cdot)\|_{\infty} \right)^2 + 2 \left(\sum_{b \in \mathcal{A}} \lambda_b \|G_b(a, \cdot)\|_{\infty} \right)^2 \\ &\leq 2 \sum_{b \in \mathcal{A}} p_b \|G_b(a, \cdot)\|_{\infty}^2 + 2 \sum_{b \in \mathcal{A}} \lambda_b \|G_b(a, \cdot)\|_{\infty}^2 \end{aligned}$$

□

In Lemma C.1 the indices b and c range over a set $\mathcal{A}$, which we take to be $[k]$ for globally observable games and Π for locally observable games. In the globally observable case the behavioral policy places mass on every action, so the sums must run over all of $[k]$; in the locally observable case both policies lie in $\Delta(\Pi)$ and the witness vanishes outside Π , so the restriction is without loss.

Lemma 3.1 (Guarantee for Alg. 2 for globally observable games). Let $\mathcal{G} = (\mathbf{L}, \Phi)$ be a globally observable partial monitoring game in the sense of Definition 2.2, and set $\beta := \max_{(a,b) \in [k] \times [k]} \|w_{(a,b)}\|_{\infty}$.

We have for any $\hat{q} \in \Delta(d)$,

$$\mathcal{V}_\gamma(\hat{q}, p) \leq \max_{\lambda \in \Delta(\Pi)} p^T \mathbf{L} \hat{q} - \lambda^T \mathbf{L} \hat{q} + \frac{16\beta^2}{\gamma} \sum_{c=1}^k \frac{1}{p_c}.$$

Thus, for every $\gamma > 16k^2\beta^2$, and every $\hat{q} \in \Delta(d)$, Alg. 2 outputs p such that

$$\mathcal{V}_\gamma(\hat{q}, p) \leq O\left(k\beta\sqrt{\frac{1}{\gamma}}\right).$$

Proof. For globally observable games, we can construct *reference witness function* by simply setting, for all $b \in [k]$

$$G_b(\cdot, \cdot) = w_{b, a_0}(\cdot, \cdot) \quad (\text{for some reference action } a_0)$$

Hence, by Lemma C.1,

$$\begin{aligned} \text{Coverage}(p, \lambda) &\leq 16 \max_{(a,b) \in [k] \times [k]} \|w_{(a,b)}\|_\infty^2 \sum_{c=1}^k \frac{1}{p_c} \\ &= 16\beta^2 \sum_{c=1}^k \frac{1}{p_c}. \end{aligned}$$

This finishes the proof of the first part of the lemma.

Let $\hat{a}$ be the empirical best arm based on our current estimate $\hat{q}$. That is,

$$\hat{a} := \arg \min_{a \in \Pi} e_a^\top \mathbf{L} \hat{q}(x).$$

Set,

$$p = (1 - \epsilon)e_{\hat{a}} + \epsilon \frac{\mathbf{1}}{k},$$

so that $p_c \geq \frac{\epsilon}{k}$ for every $c \in [k]$, hence

$$\sum_{c=1}^k \frac{1}{p_c} \leq \frac{k^2}{\epsilon}.$$

$$\begin{aligned} \mathcal{V}_\gamma(\hat{q}(x), p) &= \sup_{\lambda \in \Delta(\Pi)} \left[p^\top \mathbf{L} \hat{q}(x) - \lambda^\top \mathbf{L} \hat{q}(x) + \frac{1}{\gamma} \cdot \text{Coverage}(p, \lambda) \right] \\ &\leq p^\top \mathbf{L} \hat{q}(x) - e_{\hat{a}}^\top \mathbf{L} \hat{q}(x) + \frac{1}{\gamma} 16\beta^2 \sum_{c=1}^k \frac{1}{p_c} \\ &\leq \left((1 - \epsilon)e_{\hat{a}} + \epsilon \frac{\mathbf{1}}{k} \right)^\top \mathbf{L} \hat{q}(x) - e_{\hat{a}}^\top \mathbf{L} \hat{q}(x) + \frac{16k^2\beta^2}{\gamma\epsilon} \\ &\leq \epsilon \frac{1}{k} \mathbf{1}^\top \mathbf{L} \hat{q}(x) + \frac{16k^2\beta^2}{\gamma\epsilon} \\ &\leq \alpha + \frac{16k^2\beta^2}{\gamma\epsilon} \quad (\mathbf{L} \in [0, 1]^{k \times d}) \\ &= O\left(k\beta\sqrt{\frac{1}{\gamma}}\right). \quad (\text{by choosing } \epsilon = k\beta\sqrt{\frac{1}{\gamma}} < 1) \end{aligned}$$

This finishes the proof of the second part of the lemma. $\square$

C.3 PROOFS FOR ALGORITHM 3 (SQUARECB.PMSIDE)

We can choose a spanning in-tree over the neighborhood graph $(\Pi, \mathcal{E})$. That is, assign an arbitrary action $\text{root}_{\mathcal{T}}$ as the root and choose an in-tree $\mathcal{T}$ such that, there is a unique directed $\text{path}_{\mathcal{T}}(a)$ from any action $a \in \Pi$ to $\text{root}_{\mathcal{T}} \in \Pi$ (See Figure 1). Now, define *witness reference function* $G_{\mathcal{T}} : [k] \times \Sigma \rightarrow \mathbb{R}^{\Pi} \forall a \in \Pi$ such that,

$$G_a^{\mathcal{T}}(c, \sigma) := \sum_{e \in \text{path}_{\mathcal{T}}(a)} w_e(c, \sigma), \quad a \in \Pi, \quad (10)$$

where w_e denotes the witness function associated with the edge e by local observability (Definition 2.3). Here it is noteworthy that, let $e = (e_1, e_2)$ then $w_e(c, \sigma) = 0 \forall c \notin \{e_1, e_2\}$. Therefore,

$$G_a^{\mathcal{T}}(c, \sigma) = 0 \text{ if } c \notin \text{path}_{\mathcal{T}}(a). \quad (11)$$

Therefore, the local observability condition induces a sparse $G_a^{\mathcal{T}}(\cdot, \sigma)$ function, which will become handy for us to allocate probability only for the actions populating it. Furthermore, we can also show that, $\forall a \in \Pi$

$$\sum_{c \in \Pi} G_a^{\mathcal{T}}(c, \Phi_{co}) = \sum_{c \in \Pi} \sum_{e \in \text{path}_{\mathcal{T}}(a)} w_e(c, \Phi_{co}) = \sum_{e \in \text{path}_{\mathcal{T}}(a)} \sum_{c \in \Pi} w_e(c, \Phi_{co}) = L_{ao} - L_{\text{root}_{\mathcal{T}}o} \quad (12)$$

Hence, *witness reference function* $G_{\mathcal{T}}$ satisfies the conditions of Lemma C.1. Hence,

Now we relax $\mathcal{V}_{\gamma}(\hat{q}, p)$ with sparse $G_a^{\mathcal{T}}(\cdot, \sigma)$.

Lemma 3.3. Let $\mathcal{G} = (\mathbf{L}, \Phi)$ be a locally observable partial monitoring game in the sense of Definition 2.3. Let $\mathcal{T}$ be an in-tree of the neighborhood graph and $G^{\mathcal{T}}$ be its associated *witness reference function*. For all $p \in \Delta(\Pi)$, we have

$$\begin{aligned} \mathcal{V}_{\gamma}(\hat{q}(x), p) &\leq \max_{\lambda \in \Delta(\Pi)} p^T \mathbf{L} \hat{q}(x) - \lambda^T \mathbf{L} \hat{q}(x) + \frac{8}{\gamma} \sum_{b \in \Pi} \sum_{c \in \Pi} \left(\frac{\lambda_b}{p_c} + \frac{p_b}{p_c} \right) \|G^{\mathcal{T}}(c, \cdot)_b\|_{\infty}^2 \\ &\leq \max_{\lambda \in \Delta(\Pi)} p^T \mathbf{L} \hat{q}(x) - \lambda^T \mathbf{L} \hat{q}(x) + \frac{32}{\gamma} \max_{e \in \mathcal{E}} \|w_e\|_{\infty}^2 \sum_{(b \in \Pi, c \in \Pi) : c \in \text{path}_{\mathcal{T}}(b)} \left(\frac{\lambda_b}{p_c} + \frac{p_b}{p_c} \right). \end{aligned}$$

Proof. By Lemma C.1,

$$\text{Coverage}(p, \lambda) \leq 8 \sum_{b \in \Pi} \sum_{c \in \Pi} \left(\frac{\lambda_b}{p_c} + \frac{p_b}{p_c} \right) \|G_b^{\mathcal{T}}(c, \cdot)\|_{\infty}^2.$$

This finishes the proof of the first inequality of the lemma.

Furthermore, we have

$$G_b^{\mathcal{T}}(c, \cdot) = \sum_{e \in \text{path}_{\mathcal{T}}(b)} w_e(c, \cdot),$$

Let $e = (e_1, e_2)$, then by Definition 2.3, $w_e(c, \cdot) = 0 \forall c \neq e_1, e_2$. Hence,

$$\begin{aligned} \text{Coverage}(p, \lambda) &\leq 8 \sum_{b \in \Pi} \sum_{c \in \Pi} \left(\frac{\lambda_b}{p_c} + \frac{p_b}{p_c} \right) \|G_b^{\mathcal{T}}(c, \cdot)\|_{\infty}^2 \\ &= 8 \sum_{b \in \Pi} \sum_{c \in \Pi} \left(\frac{\lambda_b}{p_c} + \frac{p_b}{p_c} \right) \left\| \sum_{e \in \text{path}_{\mathcal{T}}(b)} w_e(c, \cdot) \mathbb{1}_{\{c \in \{e_1, e_2\}\}} \right\|_{\infty}^2 \\ &= 8 \max_{e \in \mathcal{E}} \|w_e\|_{\infty}^2 \sum_{b \in \Pi} \sum_{c \in \Pi} \left(\frac{\lambda_b}{p_c} + \frac{p_b}{p_c} \right) \left\| \sum_{e \in \text{path}_{\mathcal{T}}(b)} \mathbb{1}_{\{c \in \{e_1, e_2\}\}} \right\|_{\infty}^2 \\ &= 32 \max_{e \in \mathcal{E}} \|w_e\|_{\infty}^2 \sum_{(b, c) : c \in \text{path}_{\mathcal{T}}(b)} \left(\frac{\lambda_b}{p_c} + \frac{p_b}{p_c} \right) \\ &\quad \text{(one vertex will be a part of maximum 2 edges in an in-tree path)} \\ &\leq 32 \max_{e \in \mathcal{E}} \|w_e\|_{\infty}^2 \sum_{(b \in \Pi, c \in \Pi) : c \in \text{path}_{\mathcal{T}}(b)} \left(\frac{\lambda_b}{p_c} + \frac{p_b}{p_c} \right) \end{aligned}$$

This finishes the proof of the second inequality of the lemma. $\square$

Above result is true for any in-tree with an arbitrary root. However, in general, we do not want the paths to be arbitrary: for a “bad” action c and a “good” action b (i.e., $e_c^\top L\hat{q}$ is large while $e_b^\top L\hat{q}$ is small), we would like p_c to be small and p_b to be large to ensure $p^\top L\hat{q}(x) - \lambda^\top L\hat{q}(x)$ to be small; however, if c happens to lie in $\text{path}_\mathcal{T}(b)$, that can cause the $\frac{\lambda_b}{p_c}$ term to be very large. Hence,

1. We need to choose an in-tree $\mathcal{T}$, such that the path to any action a have actions with smaller estimated losses, such that assigning higher probability for them is both informative and rewarding.
2. We need to find a mechanism to allocate probability along this in-tree $\mathcal{T}$ respecting 1), such that both $p^\top L\hat{q}(x) - \lambda^\top L\hat{q}(x)$ and $\frac{32}{\gamma} \max_{e \in \mathcal{E}} \|w_e\|_\infty^2 \sum_{(b \in \Pi, c \in \Pi): c \in \text{path}_\mathcal{T}(b)} \left(\frac{\lambda_b}{p_c} + \frac{p_b}{p_c} \right)$ are controlled.

The best solution for 1) is an $L\hat{q}$ -decreasing in-tree. We define $\mathcal{T}$ -decreasing in-tree as follows,

Definition C.1 ($\mathcal{T}$ -decreasing in-tree [Lattimore & Szepesvári, 2020]). Let $\mathcal{T}$ be an in-tree over $[\Pi]$. A vector $y \in \mathbb{R}^\Pi$ is called $\mathcal{T}$ -decreasing if $y_a \geq y_b$ for all directed edge $(a, b) \in \mathcal{T}$. That is, $a \mapsto y_a$ is decreasing towards the root of $\mathcal{T}$.

Basically, an $L\hat{q}$ -decreasing in-tree $\mathcal{T}$ will have the empirical best action ($\arg \min_{a \in \Pi} e_a^\top L\hat{q}(x)$) as the root and for every directed edge $e = (a, b)$, $e_a^\top L\hat{q}(x) \geq e_b^\top L\hat{q}(x)$. Such a $\mathcal{T}$ is guaranteed to exist by Lattimore & Szepesvári [2020] (Lemma 19).

For 2), we observe that, inverse-gap-weighting (IGW), by construction, assigns probability in an increasing fashion toward the root. Now we provide the proof for Lemma 3.2.

Lemma 3.2 (Guarantee for Alg. 3 for locally observable games). Let $\mathcal{G} = (L, \Phi)$ be a locally observable partial monitoring game in the sense of Definition 2.3. For every $\gamma > 0$, and every $\hat{q} \in \Delta(d)$, Alg. 2 run with $\kappa = \frac{\gamma}{8km^2}$ and $\mu = k$, outputs p such that

$$\mathcal{V}_\gamma(\hat{q}, p) \leq O\left(\frac{m^2 k^2}{\gamma}\right).$$

Proof. Recall that we define in Alg. 3 that, $\hat{b} = \arg \min_{a \in \Pi} e_a^\top L\hat{q}(x)$ and $\forall a \in \Pi \setminus \{\hat{b}\}$, $p(a) = \frac{1}{\mu + \kappa(e_a - e_{\hat{b}})^\top L\hat{q}(x)}$ and $p(\hat{b}) = 1 - \sum_{a \in \Pi \setminus \{\hat{b}\}} p(a)$.

We choose $\mathcal{T}$ to be $L\hat{q}(x)$ -decreasing in-tree in the neighborhood graph; we have that for any (b, c) such that $c \in \text{path}_\mathcal{T}(b)$, $p_c \geq p_b$. This is true because for any $c \in \text{path}(b)$ of $L\hat{q}(x)$ -decreasing in-tree, estimated loss of c is lower than estimated loss of b . Hence, by construction of IGW based policy p , $p_c \geq p_b$.

By Lemma 3.3 we have,

$$\text{Coverage}(p, \lambda) \leq 32 \max_{e \in \mathcal{E}} \|w_e\|_\infty^2 \sum_{(b, c): c \in \text{path}_\mathcal{T}(b)} \left(\frac{\lambda_b}{p_c} + \frac{p_b}{p_c} \right).$$

Since for every c in $\text{path}_\mathcal{T}(b)$, $p_c \geq p_b$

$$\begin{aligned} \text{Coverage}(p, \lambda) &\leq 32 \max_{e \in \mathcal{E}} \|w_e\|_\infty^2 \sum_{b \in \Pi} \left(\frac{\lambda_b}{p_b} + \frac{p_b}{p_b} \right) \\ &\leq 32 \max_{e \in \mathcal{E}} \|w_e\|_\infty^2 k^2 + 32 \max_{e \in \mathcal{E}} \|w_e\|_\infty^2 k \sum_{b \in \Pi} \frac{\lambda_b}{p_b} \end{aligned}$$

Moreover,

$$\sup_{\lambda \in \Delta(\Pi)} \left\{ p^\top L\hat{q}(x) - \lambda^\top L\hat{q}(x) + \frac{1}{\gamma} \cdot 32 \max_{e \in \mathcal{E}} \|w_e\|_\infty^2 k \sum_{b \in \Pi} \frac{\lambda_b}{p_b} \right\}$$

$$\leq \sup_{\lambda \in \Delta(\Pi)} \left[\sum_{b \in \Pi} (p_b - \lambda_b) \mathbf{L}_{b,:} \hat{q}(x) + \frac{1}{\gamma} \cdot 32 \max_{e \in \mathcal{E}} \|w_e\|_\infty^2 k \sum_{b \in \Pi} \lambda_b \left(\mu + \kappa (e_b - e_{\hat{b}})^\top \mathbf{L} \hat{q}(x) \right) \right]$$

In the above inequality, we use the observation that $p_{\hat{b}}$ is lower bounded by $\frac{1}{k}$. This is true, since $\mu = k$.

$$\begin{aligned} & \sup_{\lambda \in \Delta(\Pi)} \left\{ p^\top \mathbf{L} \hat{q}(x) - \lambda^\top \mathbf{L} \hat{q}(x) + \frac{1}{\gamma} \cdot 32 \max_{e \in \mathcal{E}} \|w_e\|_\infty^2 k \sum_{b \in \Pi} \frac{\lambda_b}{p_b} \right\} \\ & \leq \sup_{\lambda \in \Delta(\Pi)} \left[\sum_{b \in \Pi} (p_b - \lambda_b) \mathbf{L}_{b,:} \hat{q}(x) + \frac{1}{\gamma} \cdot 32 \max_{e \in \mathcal{E}} \|w_e\|_\infty^2 k^2 + \sum_{b \in \Pi} \lambda_b (e_b - e_{\hat{b}})^\top \mathbf{L} \hat{q}(x) \right] \\ & \quad (\mu = k, \kappa = \frac{\gamma}{32 \max_{e \in \mathcal{E}} \|w_e\|_\infty^2 k}) \end{aligned}$$

Furthermore,

$$\begin{aligned} & \sup_{\lambda \in \Delta(\Pi)} \left[\sum_{b \in \Pi} (p_b - \lambda_b) \mathbf{L}_{b,:} \hat{q}(x) + \sum_{b \in \Pi} \lambda_b (e_b - e_{\hat{b}})^\top \mathbf{L} \hat{q}(x) \right] \\ & = \sum_{b \in \Pi} p_b \cdot e_b^\top \mathbf{L} \hat{q}(x) - e_{\hat{b}}^\top \mathbf{L} \hat{q}(x) \\ & = \sum_{b \in \Pi \setminus \{\hat{b}\}} p_b \cdot e_b^\top \mathbf{L} \hat{q}(x) - e_{\hat{b}}^\top \mathbf{L} \hat{q}(x) \\ & = \sum_{b \in \Pi \setminus \{\hat{b}\}} \frac{1}{\mu + \kappa (e_a - e_{\hat{b}})^\top \mathbf{L} \hat{q}(x)} \cdot e_b^\top \mathbf{L} \hat{q}(x) - e_{\hat{b}}^\top \mathbf{L} \hat{q}(x) \\ & \leq \frac{k}{\kappa} = \frac{1}{\gamma} \cdot 32 \max_{e \in \mathcal{E}} \|w_e\|_\infty^2 k^2 \end{aligned}$$

Hence, $\forall \hat{q} \in \Delta(d)$

$$\mathcal{V}_\gamma(\hat{q}, p) \leq \frac{1}{\gamma} \cdot 96 \max_{e \in \mathcal{E}} \|w_e\|_\infty^2 k^2.$$

Furthermore, revisiting Definition 2.3,

$$\begin{aligned} \mathbf{L}_{ao} - \mathbf{L}_{bo} &= \sum_{c \in \{a, b\}} w_e(c, \Phi_{co}), \quad \text{for all } o \in [d]. \\ &= w_e(a, \Phi_{co}) + w_e(b, \Phi_{co}). \end{aligned}$$

By Lemma G.3, one can always find, a witness function such that, $\|w_e\|_\infty \leq \frac{m}{2}$, Hence, $\forall \hat{q} \in \Delta(d)$

$$\mathcal{V}_\gamma(\hat{q}, p) = O\left(\frac{m^2 k^2}{\gamma}\right).$$

This completes the proof. $\square$

C.4 NON-DEGENERACY ASSUMPTION

Definition 2.3 is applicable only when the game is non-degenerate. When there is degenerate action in $[k]$, the definition for local observability becomes,

Definition C.2 (Local observability, general form; Bartók et al. [2012b]). A globally observable partial monitoring game is *locally observable* if for every edge $e = (a, b) \in \mathcal{E}$ of the neighborhood

graph the witness function w_e of Definition 2.2 may be chosen supported on

$$N_{a,b}^+ := \{c \in [k] : \mathcal{C}_a \cap \mathcal{C}_b \subseteq \mathcal{C}_c\},$$

that is, $w_e(c, \cdot) \equiv 0$ for all $c \notin N_{a,b}^+$, so that

$$L_{ao} - L_{bo} = \sum_{c \in N_{a,b}^+} w_e(c, \Phi_{co}), \quad \text{for all } o \in [d].$$

The set $N_{a,b}^+$ contains the two endpoints, but may also contain degenerate actions whose cells lie inside the shared face $\mathcal{C}_a \cap \mathcal{C}_b$. For $d = 3$, for instance, a degenerate action with $\dim(\mathcal{C}_c) = 1$ is optimal only along a line segment, whereas a Pareto optimal action is optimal on a two-dimensional region, yet such an action can still be a necessary source of information.

Two parts of our development break under this more general condition. First, the exploration rule must assign probability to the degenerate actions as well, which requires a modification of SQUARECB.PMSIDE. Second, Lemma G.3 no longer applies and must be replaced by Proposition 21 of Lattimore & Szepesvári [2020] (see Lemma G.4 for restatement), which introduces an exponential factor in d . The modification to the algorithm is straightforward, and we conjecture that SQUARECB.PMSIDE continues to guarantee $O(\sqrt{T})$ regret, but the analysis requires care. We leave this extension to future work.

D CONTEXTUAL PARTIAL MONITORING: REDUCTION TO OFFLINE REGRESSION

For definitions and examples of offline outcome estimation oracles, Please see Section B.2.

D.1 OE2D.PMSIDE

The bounds on $\mathcal{V}_\gamma(\hat{q}(x), p(x))$ established for globally observable games (Lemma 3.1) and for locally observable games (Lemma 3.2), under the exploration strategies ϵ -GREEDY.PMSIDE and SQUARECB.PMSIDE respectively, remain valid in the offline setting, since both hold for every $\hat{q} \in \Delta(d)$ rather than for a particular estimate.

What does not carry over is the regret decomposition of Theorem 3.1. Inspecting the guarantee of Assumption B.1, an offline oracle's Hellinger bound makes two demands of the sampling policy: it must be held fixed over the rounds to which the bound is applied, and it must be the policy that collected the data from which $\hat{q}$ was estimated.

The first is obtained by freezing the policy within each epoch (Line 8). The second is where coverage does the work. Because $\text{Coverage}(p, \lambda)$ is a supremum over pairs of models in the class and never references $\hat{q}$, a policy with small coverage distinguishes *any* two members of $\mathcal{Q}$, in particular $\hat{q}$ and q^* , whichever estimate the oracle happens to return. Lemma D.1 makes this precise.

The DEC of Foster et al. [2021] lacks this property: its exploration term $-\gamma \cdot \mathbb{E}_{a \sim p} \left[D_H^2(\mathcal{S}_a \hat{q}(x) \mid \mathcal{S}_a q) \right]$ is anchored at $\hat{q}$. In an epoch-based algorithm (Lines 7–14), the estimation error for a given $\hat{q}$ therefore cannot be controlled using a policy from a different epoch, which played no part in producing it.

Now we present our algorithm OE2D.PMSIDE.

Furthermore, we can employ Algorithm 2 for globally observable games, such that for every $\gamma_\ell > 16k^2\beta^2$, and every $\hat{q} \in \Delta(d)$, Alg. 2 outputs p such that

$$\mathcal{V}_{\gamma_\ell}(\hat{q}, p) \leq O\left(k\beta\sqrt{\frac{1}{\gamma_\ell}}\right). \quad (\text{Lemma 3.1})$$

and for locally observable games, we can employ Algorithm 3, such that For every $\gamma_\ell > 0$, and every $\hat{q} \in \Delta(d)$, Alg. 2 run with $\kappa = \frac{\gamma}{8k^2m^2}$ and $\mu = k$, outputs p such that

$$\mathcal{V}_{\gamma_\ell}(\hat{q}, p) \leq O\left(\frac{m^2k^3}{\gamma_\ell}\right). \quad (\text{Lemma 3.2})$$

Algorithm 4 OE2D.PMSIDE

```

1: parameters:
   Offline estimation oracle  $\text{Alg}_{\text{OFF}}$ 
   Exploration parameters  $\{\gamma_\ell\}_{\ell=1}^L$ 
   Error rates  $\{\delta_\ell\}_{\ell=1}^L$ 
   Epoch schedule  $0 = \tau_0 < \tau_1 < \dots < \tau_L$ 
2: for  $t = 1$  to  $\tau_1$  do
3:   Observe the context  $x_t$ 
4:   Find  $p_t$  that approximately minimizes,
       
$$\sup_{\lambda \in \Delta(\Pi)} \text{Coverage}(p, \lambda).$$

5:   Choose action  $A_t \sim p_t$  and observe feedback  $\sigma_t$ 
6: end for
7: for  $\ell = 2, 3, \dots, L$  do
8:   Compute estimate  $\hat{q}_\ell = \text{Alg}_{\text{OFF}}\left(\{A_s, x_s, \sigma_s\}_{s=\tau_{\ell-2}+1}^{\tau_{\ell-1}}\right)$ 
9:   for  $t = \tau_{\ell-1} + 1$  to  $\tau_\ell$  do
10:    Receive context  $x_t$ 
11:    Find  $p_t$  that approximately minimizes
       
$$\mathcal{V}_{\gamma_\ell}(\hat{q}_\ell(x_t), p) := \sup_{\lambda \in \Delta(\Pi)} \underbrace{p^\top \mathbf{L} \hat{q}_\ell(x_t) - \lambda^\top \mathbf{L} \hat{q}_\ell(x_t)}_{\text{exploit}} + \frac{1}{\gamma_\ell} \cdot \underbrace{\text{Coverage}(p, \lambda)}_{\text{explore}}.$$

12:    Choose action  $A_t \sim p_t$  and observe feedback  $\sigma_t$ .
13:   end for
14: end for

```

So in both cases, we have algorithms to find policies that establishes an upper bound for our objective function. Let us denote it by V_ℓ . So,

$$V_\ell = O\left(k\beta\sqrt{\frac{1}{\gamma_\ell}}\right) \quad (\text{for globally observable games})$$

$$V_\ell = O\left(\frac{m^2 k^3}{\gamma_\ell}\right) \quad (\text{for locally observable games})$$

Furthermore, within a single epoch ℓ for all round $t \in [\tau_{\ell-1} + 1, \tau_\ell]$, the estimator does not change ($\hat{q}_\ell(\cdot)$), only the context x_t changes. So, within an epoch ℓ ,

$$x_{t_1} = x_{t_2} \implies \hat{q}_\ell(x_{t_1}) = \hat{q}_\ell(x_{t_2}) \implies p_{t_1} = p_{t_2} \quad \forall t_1, t_2 \in [\tau_{\ell-1} + 1, \tau_\ell]$$

So, we can always think that within an epoch, every p_t is sampled from a fixed distribution $p_\ell(\cdot | x_t)$, that is, we can denote,

$$p_t \leftarrow p_\ell(\cdot | x_t) \quad \forall t \in [\tau_{\ell-1} + 1, \tau_\ell].$$

D.2 REGRET ANALYSIS

Next, for readability and reduce the notations clutter we define the following,

Definition D.1. Define:

- $\mathcal{L}^*(\lambda | x) = \lambda^\top \mathbf{L} q^*(x)$: The loss incurred by policy λ under true outcome distribution q^* for a context x .
- $\widehat{\mathcal{L}}_\ell(\lambda | x) = \lambda^\top \mathbf{L} \hat{q}_\ell(x)$: The loss incurred by policy λ under estimated distribution $\hat{q}_\ell$ for a context x .
- $\lambda^* = \arg \min_{\lambda \in \Delta(\Pi)} \mathcal{L}^*(\lambda | x)$: The true best policy for context x .
- $\hat{\lambda}_\ell = \arg \min_{\lambda \in \Delta(\Pi)} \widehat{\mathcal{L}}_\ell(\lambda | x)$: The estimated best policy for context x .
- $\text{Reg}(\lambda | x) = \mathcal{L}^*(\lambda | x) - \mathcal{L}^*(\lambda^* | x)$: The regret incurred by policy λ under true outcome distribution q^* for a context x .

- $\widehat{\text{Reg}}_\ell(\lambda \mid x) = \widehat{\mathcal{L}}_\ell(\lambda \mid x) - \widehat{\mathcal{L}}_\ell(\hat{\lambda} \mid x)$: The regret incurred by policy λ under estimated distribution $\hat{q}_\ell$ for a context x .
- $C(p_\ell, \lambda) = \text{Coverage}(p_\ell, \lambda)$
- $E_\ell = \mathbb{E}_{a \sim p_{\ell-1}} \left[D_H^2(\mathcal{S}_a \hat{q}_\ell(x) \mid \mathcal{S}_a q^*(x)) \right]$

In all the above definitions, we might drop $(\cdot \mid x)$, when it is evident from the context. For example, we might denote $\mathcal{L}^*(\lambda \mid x) \rightarrow \mathcal{L}^*(\lambda)$.

Next we present two important properties of our policies and estimators.

Certification:

This property is a direct consequence of our algorithms finding the policies which bounds our objective function by V_ℓ .

$\forall x \in \mathcal{X}$ and $\forall \ell \geq 1$,

$$\mathcal{V}_{\gamma_\ell}(\hat{q}_\ell(x), p_\ell) = \sup_{\lambda \in \Delta(\Pi)} p_\ell^\top \mathbf{L} \hat{q}_\ell(x) - \lambda^\top \mathbf{L} \hat{q}_\ell(x) + \frac{1}{\gamma_\ell} \cdot \text{Coverage}(p_\ell, \lambda) \leq V_\ell$$

Hence,

$\forall x \in \mathcal{X}, \forall \ell \geq 1$ and, $\forall \lambda \in \Pi$

$$\begin{aligned} p_\ell^\top \mathbf{L} \hat{q}_\ell(x) - \lambda^\top \mathbf{L} \hat{q}_\ell(x) + \frac{1}{\gamma_\ell} \cdot \text{Coverage}(p_\ell, \lambda) &\leq V_\ell \\ \implies \widehat{\mathcal{L}}_\ell(p_\ell \mid x) - \widehat{\mathcal{L}}_\ell(\lambda \mid x) + \frac{1}{\gamma_\ell} \cdot C(p_\ell, \lambda) &\leq V_\ell \end{aligned}$$

Since, this is true for all $x \in \mathcal{X}$, we take the liberty to drop it,

$$\widehat{\mathcal{L}}_\ell(p_\ell) - \widehat{\mathcal{L}}_\ell(\lambda) + \frac{1}{\gamma_\ell} \cdot C(p_\ell, \lambda) \leq V_\ell. \quad (\text{certification})$$

Off-policy evaluation (OPE):

Lemma D.1 (Off-policy evaluation lemma). With all the definitions in Definition D.1, $\forall \lambda, \forall \ell \geq 2$, and $\forall x$,

$$\left| \widehat{\mathcal{L}}_\ell(\lambda - p_{\ell-1} \mid x) - \mathcal{L}^*(\lambda - p_{\ell-1} \mid x) \right| \leq \frac{1}{6\gamma_\ell} \cdot C(p_{\ell-1}, \lambda) + \frac{3\gamma_\ell}{2} E_\ell. \quad (\text{Off-policy evaluation})$$

Proof.

$$\begin{aligned} &\left(\widehat{\mathcal{L}}_\ell(\lambda - p_{\ell-1} \mid x) - \mathcal{L}^*(\lambda - p_{\ell-1} \mid x) \right)^2 \\ &= \left((\lambda - p_{\ell-1})^T \mathbf{L} \hat{q}_\ell(x) - (\lambda - p_{\ell-1})^T \mathbf{L} q^*(x) \right)^2 \\ &= \frac{\left((\lambda - p_{\ell-1})^T \mathbf{L} (\hat{q}_\ell(x) - q^*(x)) \right)^2}{\mathbb{E}_{a \sim p_{\ell-1}} \left[D_H^2(\mathcal{S}_a \hat{q}_\ell(x) \mid \mathcal{S}_a q^*(x)) \right]} \cdot \mathbb{E}_{a \sim p_{\ell-1}} \left[D_H^2(\mathcal{S}_a \hat{q}_\ell(x) \mid \mathcal{S}_a q^*(x)) \right] \\ &\leq \text{Coverage}(p_{\ell-1}, \lambda) \cdot \mathbb{E}_{a \sim p_{\ell-1}} \left[D_H^2(\mathcal{S}_a \hat{q}_\ell(x) \mid \mathcal{S}_a q^*(x)) \right] \\ &\left| \widehat{\mathcal{L}}_\ell(\lambda - p_{\ell-1} \mid x) - \mathcal{L}^*(\lambda - p_{\ell-1} \mid x) \right| \\ &\leq \sqrt{\text{Coverage}(p_{\ell-1}, \lambda) \cdot \mathbb{E}_{a \sim p_{\ell-1}} \left[D_H^2(\mathcal{S}_a \hat{q}_\ell(x) \mid \mathcal{S}_a q^*(x)) \right]} \\ &\leq \frac{1}{6\gamma_\ell} \cdot \text{Coverage}(p_{\ell-1}, \lambda) + \frac{3\gamma_\ell}{2} \mathbb{E}_{a \sim p_{\ell-1}} \left[D_H^2(\mathcal{S}_a \hat{q}_\ell(x) \mid \mathcal{S}_a q^*(x)) \right] \quad (\text{AM-GM}) \end{aligned}$$

$$= \frac{1}{6\gamma_\ell} \cdot C(p_{\ell-1}, \lambda) + \frac{3\gamma_\ell}{2} E_\ell.$$

□

The second term is the generalization error of the offline oracle: $\hat{q}_\ell$ is estimated from data collected in the previous epoch under $p_{\ell-1}$, so Assumption B.1 (Hellinger boundedness) applies to it directly. This coupling between the estimate and the policy that produced it is what makes the offline analysis more delicate than the online one, where the estimator is updated as the data arrive.

The first term involves no estimate at all: $\text{Coverage}(p_{\ell-1}, \lambda)$ depends only on the model class, on λ and on $p_{\ell-1}$. A policy that makes it small therefore controls the evaluation error for every $\hat{q} \in \text{co}(\mathcal{Q}_x)$ at once, in particular for whichever estimate the oracle returns in epoch ℓ .

Now we present an abstract regret theorem about epoch-based offline oracle-efficient interactive learning algorithm from Qin & Zhang [2026a].

Theorem D.1 (Generalization of regret concentration theorem from Qin & Zhang [2026a]). Let Λ be a set of policies, $\mathcal{L}(\cdot | x) : \Lambda \rightarrow [0, 1]$ a value function and $\mathcal{L}^*(\cdot | x)$ and $\widehat{\mathcal{L}}_\ell(\cdot | x)$, $\ell \in [L]$ are value functions measured under the true model and estimated models respectively. $\lambda^* \in \Lambda$ and $\hat{\lambda}_\ell \in \Lambda$ are respectively the optimal policies under respective true and estimated models. Let $C_\ell : \Lambda \times \Lambda \rightarrow \mathbb{R}_+$ be coverage maps, $p_\ell \in \Lambda$ are decisions, $0 < \gamma_1 \leq \gamma_2 \leq \dots \leq \gamma_L$ learning rates, and $V_\ell, E_\ell \geq 0$ numbers such that,

(C1) (certificate) for every $\ell \geq 1$ and every $\lambda \in \Lambda$,

$$\widehat{\mathcal{L}}_\ell(p_\ell | x) - \widehat{\mathcal{L}}_\ell(\lambda | x) + \frac{1}{\gamma_\ell} C_\ell(p_\ell, \lambda) \leq V_\ell.$$

(C2) (off-policy evaluation) for every $\ell \geq 2$ and every $\lambda \in \Lambda$,

$$\left| \widehat{\mathcal{L}}_\ell(\lambda - p_{\ell-1} | x) - \mathcal{L}^*(\lambda - p_{\ell-1} | x) \right| \leq \frac{1}{6\gamma_\ell} C_{\ell-1}(p_{\ell-1}, \lambda) + \frac{3\gamma_\ell}{2} E_\ell.$$

Define $G_1 := 1$ and, for $\ell \geq 2$,

$$\gamma_\ell G_\ell := \gamma_{\ell-1} G_{\ell-1} + \gamma_{\ell-1} V_{\ell-1} + 9\gamma_\ell^2 E_\ell, \quad \text{i.e.} \quad G_\ell = \frac{1}{\gamma_\ell} \left[\gamma_1 + \sum_{i=2}^{\ell} (\gamma_{i-1} V_{i-1} + 9\gamma_i^2 E_i) \right].$$

Then, for every $\ell \in [L]$ and every $\lambda \in \Lambda$,

$$\text{Reg}(\lambda | x) \leq 2\widehat{\text{Reg}}_\ell(\lambda | x) + G_\ell \quad \text{and} \quad \widehat{\text{Reg}}_\ell(\lambda | x) \leq 2\text{Reg}(\lambda | x) + G_\ell.$$

Consequently $\text{Reg}(p_\ell | x) \leq 2V_\ell + G_\ell$ for every ℓ , and

$$\sum_{\ell=1}^L (\tau_\ell - \tau_{\ell-1}) \text{Reg}(p_\ell | x) \leq \tau_1 + \sum_{\ell=2}^L \frac{\tau_\ell - \tau_{\ell-1}}{\gamma_\ell} \left[2\gamma_\ell V_\ell + \gamma_1 + \sum_{i=2}^{\ell} (\gamma_{i-1} V_{i-1} + 9\gamma_i^2 E_i) \right].$$

Proof is deferred to Appendix G. We can apply this theorem directly and obtain a per-context regret bound. Now we state the main theorem that bounds the regret with high probability for any epoch schedule.

Theorem D.2 (Regret of OE2D.PMSIDE). Suppose Alg_{Off} has a Hellinger regret bound $\overline{\text{Reg}}_{\text{Off}}^{\text{H}}(\mathcal{Q}, T, \delta)$. For any non-degenerate globally observable partial monitoring game without duplicate actions, assume that $x_t \stackrel{i.i.d.}{\sim} \mathcal{D}_x$, then for any $\delta \in (0, 1)$, OE2D.PMSIDE run with

1. exploration rule Algorithm 2 for globally observable game (with $\gamma_\ell > 16k^2\beta^2 \forall \ell \in [L]$).
2. exploration rule Algorithm 3 for locally observable game (with $\kappa = \frac{\gamma_\ell}{8k^2m^2} \forall \ell \in [L]$ and $\mu = k$)
3. exploration parameters such that, $0 < \gamma_1 \leq \gamma_2 \leq \dots \leq \gamma_L$

satisfies,

$$\widetilde{\text{Reg}}_T \leq L \cdot \tau_1 + 2L \max_{\ell \in [L]} \frac{\tau_\ell - \tau_{\ell-1}}{\gamma_\ell} \left(2\gamma_\ell V_\ell + \gamma_1 + \sum_{i=2}^{\ell} \left(\gamma_{i-1} V_{i-1} + 9\gamma_i^2 \overline{\text{Reg}}_{\text{Off}}^{\text{H}}(\mathcal{Q}, \tau_i - \tau_{i-1}, \delta_i) \right) \right).$$

where,

- $\widetilde{\text{Reg}}_T = \mathbb{E}_{x_1, \dots, x_T \stackrel{i.i.d.}{\sim} \mathcal{D}_x} [\text{Reg}_T]$
- $V_\ell = O\left(k\beta\sqrt{\frac{1}{\gamma_\ell}}\right)$ for globally observable games
- $V_\ell = O\left(\frac{m^2 k^3}{\gamma_\ell}\right)$ for locally observable games

Proof. From Theorem D.1, the regret per-context is bounded by,

$$\begin{aligned} \sum_{\ell=1}^L (\tau_\ell - \tau_{\ell-1}) \text{Reg}(p_\ell | x) &\leq \tau_1 + \sum_{\ell=2}^L (\tau_\ell - \tau_{\ell-1}) (2V_\ell + G_\ell) & (\text{Reg}(p_1 | x) \leq 1) \\ &= \tau_1 + \sum_{\ell=2}^L \frac{\tau_\ell - \tau_{\ell-1}}{\gamma_\ell} \left[2\gamma_\ell V_\ell + \gamma_1 + \sum_{i=2}^{\ell} (\gamma_{i-1} V_{i-1} + 9\gamma_i^2 E_i) \right]. \end{aligned}$$

Let us denote,

$$\widetilde{\text{Reg}}_T = \mathbb{E}_{x \stackrel{i.i.d.}{\sim} \mathcal{D}_x} [\text{Reg}_T]$$

Here, note that, even though the randomness w.r.t to contexts are removed, $\widetilde{\text{Reg}}_T$, is still a random variable in terms of policy. Hence,

$$\begin{aligned} \widetilde{\text{Reg}}_T &= \mathbb{E}_{x \stackrel{i.i.d.}{\sim} \mathcal{D}_x} [\text{Reg}_T] = \mathbb{E} \left[\sum_{\ell=1}^L \sum_{t=\tau_{\ell-1}}^{\tau_\ell} \text{Reg}(p_\ell | x_t) \right] \\ &\leq \mathbb{E} \left[\sum_{\ell=1}^L \tau_1 + \sum_{\ell=2}^L \frac{\tau_\ell - \tau_{\ell-1}}{\gamma_\ell} \left(2\gamma_\ell V_\ell + \gamma_1 + \sum_{i=2}^{\ell} (\gamma_{i-1} V_{i-1} + 9\gamma_i^2 E_i) \right) \right] \\ &\leq L \cdot \tau_1 + 2L \max_{\ell \in [L]} \frac{\tau_\ell - \tau_{\ell-1}}{\gamma_\ell} \left(2\gamma_\ell V_\ell + \gamma_1 + \sum_{i=2}^{\ell} \left(\gamma_{i-1} V_{i-1} + 9\gamma_i^2 \mathbb{E}_{x \stackrel{i.i.d.}{\sim} \mathcal{D}_x} [E_i] \right) \right). \end{aligned}$$

If the contexts are i.i.d such that, $x_t \stackrel{i.i.d.}{\sim} \mathcal{D}_x$, then,

$$\begin{aligned} \widetilde{\text{Reg}}_T &\leq L \cdot \tau_1 \\ &\quad + 2L \max_{\ell \in [L]} \frac{\tau_\ell - \tau_{\ell-1}}{\gamma_\ell} \left(2\gamma_\ell V_\ell + \gamma_1 + \sum_{i=2}^{\ell} \left(\gamma_{i-1} V_{i-1} + 9\gamma_i^2 \mathbb{E}_{x \stackrel{i.i.d.}{\sim} \mathcal{D}_x} \mathbb{E}_{a \sim p_{\ell-1}} \left[D_H^2(\mathbf{S}_a \hat{q}_\ell(x) | \mathbf{S}_a q^*(x)) \right] \right) \right) \\ &\leq L \cdot \tau_1 + 2L \max_{\ell \in [L]} \frac{\tau_\ell - \tau_{\ell-1}}{\gamma_\ell} \left(2\gamma_\ell V_\ell + \gamma_1 + \sum_{i=2}^{\ell} \left(\gamma_{i-1} D_{i-1} + 9\gamma_i^2 \frac{\text{Reg}_{\text{Off}}^{\text{H}}(\mathcal{Q}, \tau_i - \tau_{i-1})}{\tau_i - \tau_{i-1}} \right) \right). \end{aligned}$$

We know that, with probability of at least $1 - \delta'$,

$$\text{Reg}_{\text{Off}}^{\text{H}}(\mathcal{Q}, \tau_i - \tau_{i-1}) \leq \overline{\text{Reg}}_{\text{Off}}^{\text{H}}(\mathcal{Q}, \tau_i - \tau_{i-1}, \delta)$$

Set $\delta_\ell = \frac{\delta}{L}$, with a probability at least $1 - \delta$,

$$\widetilde{\text{Reg}}_T \leq L \cdot \tau_1 + 2L \max_{\ell \in [L]} \frac{\tau_\ell - \tau_{\ell-1}}{\gamma_\ell} \left(2\gamma_\ell V_\ell + \gamma_1 + \sum_{i=2}^{\ell} \left(\gamma_{i-1} V_{i-1} + 9\gamma_i^2 \frac{\overline{\text{Reg}}_{\text{Off}}^{\text{H}}(\mathcal{Q}, \tau_i - \tau_{i-1}, \delta_i)}{\tau_i - \tau_{i-1}} \right) \right).$$

□

The final regret depends on epoch schedule. We choose a doubling epoch schedule.

Corollary 3. Suppose Alg_{Off} has a Hellinger regret bound $\overline{\text{Reg}}_{\text{Off}}^{\text{H}}(\mathcal{Q}, T, \delta)$. For any non-degenerate globally observable partial monitoring game without duplicate actions, assume that $x_1, \dots, x_T \stackrel{i.i.d}{\sim} \mathcal{D}_x$, then for any $\delta \in (0, 1)$, the algorithm OE2D.PMSIDE run with,

1. exploration rule Algorithm 2 for globally observable game (with $\gamma_\ell > 16k^2\beta^2 \forall \ell \in [L]$).
2. exploration rule Algorithm 3 for locally observable game (with $\kappa = \frac{\gamma_\ell}{8k^2m^2} \forall \ell \in [L]$ and $\mu = k$)
3. doubling schedule, such that $\tau_\ell = 2^\ell$
4. exploration parameters such that, $0 < \gamma_1 \leq \gamma_2 \leq \dots \leq \gamma_L$
5. with assumption $\overline{\text{Reg}}_{\text{Off}}^{\text{H}}(\mathcal{Q}, t, \delta)$ does not depend on t . Generally this is the case.
6. with $\delta_\ell = \frac{\delta}{L}$

satisfies the following guarantees,

1. For globally observable games, with $\gamma_\ell = \left(\frac{C \cdot \tau_\ell \cdot k \beta}{\overline{\text{Reg}}_{\text{Off}}^{\text{H}}(\mathcal{Q}, \tau_\ell/2, \delta/L)} \right)^{\frac{2}{3}}$ for some constant C ,

$$\widetilde{\text{Reg}}_T \leq O\left(T^{\frac{2}{3}} \log^2(T) (k\beta)^{\frac{2}{3}} (\overline{\text{Reg}}_{\text{Off}}^{\text{H}}(\mathcal{Q}, T, \delta/\log(T)))^{\frac{1}{3}}\right)$$

2. For locally observable games, with $\gamma_\ell = \sqrt{\frac{D \cdot \tau_\ell \cdot k^3 m^2}{\overline{\text{Reg}}_{\text{Off}}^{\text{H}}(\mathcal{Q}, \tau_\ell/2, \delta_\ell)}}$ for some constant D ,

$$\widetilde{\text{Reg}}_T = O\left(\log^2(T) \sqrt{T \cdot k^3 m^2 \overline{\text{Reg}}_{\text{Off}}^{\text{H}}(\mathcal{Q}, T/2, \delta/\log(T))}\right)$$

where, $\widetilde{\text{Reg}}_T = \mathbb{E}_{x \stackrel{i.i.d}{\sim} \mathcal{D}_x} [\text{Reg}_T]$.

Proof. First of all, for doubling epoch schedule,

$$\begin{aligned} \tau_L &= O(T) \\ \tau_\ell - \tau_{\ell-1} &= \frac{\tau_\ell}{2} \\ L &= O(\log T) \end{aligned}$$

$$\begin{aligned} \widetilde{\text{Reg}}_T &\leq L \cdot \tau_1 + 2L \max_{\ell \in [L]} \frac{\tau_\ell - \tau_{\ell-1}}{\gamma_\ell} \left(2\gamma_\ell V_\ell + \gamma_1 + \sum_{i=2}^{\ell} \left(\gamma_{i-1} V_{i-1} + 9\gamma_i^2 \frac{\overline{\text{Reg}}_{\text{Off}}^{\text{H}}(\mathcal{Q}, \tau_i - \tau_{i-1}, \delta_i)}{\tau_i - \tau_{i-1}} \right) \right) \\ &= L \cdot \tau_1 + 2L \max_{\ell \in [L]} \frac{\tau_\ell}{2\gamma_\ell} \left(2\gamma_\ell V_\ell + \gamma_1 + \sum_{i=2}^{\ell} \left(\gamma_{i-1} V_{i-1} + 9\gamma_i^2 \frac{\overline{\text{Reg}}_{\text{Off}}^{\text{H}}(\mathcal{Q}, \tau_i/2, \delta_i)}{\tau_i/2} \right) \right) \\ &\leq L \cdot \tau_1 + 2L \max_{\ell \in [L]} \frac{\tau_\ell}{2\gamma_\ell} \left(\gamma_1 + \sum_{i=2}^{\ell} \gamma_i \left(2V_i + 18\gamma_i \frac{\overline{\text{Reg}}_{\text{Off}}^{\text{H}}(\mathcal{Q}, \tau_i/2, \delta_i)}{\tau_i} \right) \right) \\ &\leq L \cdot \tau_1 + 2L^2 \max_{\ell \in [L]} \frac{\tau_\ell}{2\gamma_\ell} \left(\gamma_1 + \max_{i \in [\ell]} \gamma_i \left(2V_i + 18\gamma_i \frac{\overline{\text{Reg}}_{\text{Off}}^{\text{H}}(\mathcal{Q}, \tau_i/2, \delta_i)}{\tau_i} \right) \right) \\ &= L \cdot \tau_1 + 2L^2 \max_{\ell \in [L]} \frac{\tau_\ell}{2\gamma_\ell} \left(\gamma_1 + \max_{i \in [\ell]} \frac{\gamma_i}{\tau_i} \left(2\tau_i V_i + 18\gamma_i \overline{\text{Reg}}_{\text{Off}}^{\text{H}}(\mathcal{Q}, \tau_i/2, \delta_i) \right) \right). \end{aligned}$$

For globally observable games,

$$V_i = C \cdot k\beta \sqrt{\frac{1}{\gamma_i}} \quad (\text{for some constant } C)$$

$$\widetilde{\text{Reg}}_T \leq L \cdot \tau_1 + 2L^2 \max_{\ell \in [L]} \frac{\tau_\ell}{2\gamma_\ell} \left(\gamma_1 + \max_{i \in [\ell]} \frac{\gamma_i}{\tau_i} \left(2\tau_i C \cdot k\beta \sqrt{\frac{1}{\gamma_i}} + 18\gamma_i \overline{\text{Reg}}_{\text{Off}}^H(\mathcal{Q}, \tau_i/2, \delta_i) \right) \right).$$

$$\text{set } \gamma_i = \left(\frac{C\tau_i \cdot k\beta}{\overline{\text{Reg}}_{\text{Off}}^H(\mathcal{Q}, \tau_i/2, \delta_i)} \right)^{\frac{2}{3}}$$

$$\begin{aligned} \widetilde{\text{Reg}}_T &\leq L \cdot \tau_1 + 2L^2 \max_{\ell \in [L]} \frac{\tau_\ell}{2\gamma_\ell} \left(\gamma_1 + \max_{i \in [\ell]} \frac{\gamma_i}{\tau_i} \left(2 \cdot (C \cdot \tau_i k\beta)^{\frac{2}{3}} (\overline{\text{Reg}}_{\text{Off}}^H(\mathcal{Q}, \tau_i/2, \delta_i))^{\frac{1}{3}} + 18(C \cdot \tau_i k\beta)^{\frac{2}{3}} (\overline{\text{Reg}}_{\text{Off}}^H(\mathcal{Q}, \tau_i/2, \delta_i))^{\frac{1}{3}} \right) \right) \\ &= L \cdot \tau_1 + 2L^2 \max_{\ell \in [L]} \frac{\tau_\ell}{2\gamma_\ell} \left(\gamma_1 + 20 \max_{i \in [\ell]} \frac{\gamma_i}{\tau_i} (C \cdot \tau_i k\beta)^{\frac{2}{3}} (\overline{\text{Reg}}_{\text{Off}}^H(\mathcal{Q}, \tau_i/2, \delta_i))^{\frac{1}{3}} \right) \\ &= L \cdot \tau_1 + 2L^2 \max_{\ell \in [L]} \frac{\tau_\ell}{2\gamma_\ell} \left(\gamma_1 + 20 \max_{i \in [\ell]} \frac{1}{\tau_i} (C \cdot \tau_i k\beta)^{\frac{4}{3}} (\overline{\text{Reg}}_{\text{Off}}^H(\mathcal{Q}, \tau_i/2, \delta_i))^{-\frac{1}{3}} \right) \\ &\leq L \cdot \tau_1 + 2L^2 \max_{\ell \in [L]} \frac{\tau_\ell}{2\gamma_\ell} \left(\gamma_1 + 20\tau_\ell^{\frac{1}{3}} (C \cdot k\beta)^{\frac{4}{3}} (\max_{i \in [\ell]} \overline{\text{Reg}}_{\text{Off}}^H(\mathcal{Q}, \tau_i/2, \delta_i))^{-\frac{1}{3}} \right) \\ &= L \cdot \tau_1 + L^2 \max_{\ell \in [L]} \frac{(\overline{\text{Reg}}_{\text{Off}}^H(\mathcal{Q}, \tau_\ell/2, \delta_\ell))^{\frac{2}{3}} \tau_\ell^{\frac{1}{3}}}{(C \cdot k\beta)^{\frac{2}{3}}} \left(\gamma_1 + 20\tau_\ell^{\frac{1}{3}} (C \cdot k\beta)^{\frac{4}{3}} (\max_{i \in [\ell]} \overline{\text{Reg}}_{\text{Off}}^H(\mathcal{Q}, \tau_i/2, \delta_i))^{-\frac{1}{3}} \right) \\ &= O \left(L^2 (C \cdot k\beta)^{\frac{2}{3}} \max_{\ell \in [L]} \tau_\ell^{\frac{2}{3}} \frac{(\overline{\text{Reg}}_{\text{Off}}^H(\mathcal{Q}, \tau_\ell/2, \delta_\ell))^{\frac{2}{3}}}{\min_{\ell \in [L]} (\overline{\text{Reg}}_{\text{Off}}^H(\mathcal{Q}, \tau_\ell/2, \delta_\ell))^{\frac{1}{3}}} \right) \\ &= O \left(T^{\frac{2}{3}} \log^2(T) (k\beta)^{\frac{2}{3}} (\overline{\text{Reg}}_{\text{Off}}^H(\mathcal{Q}, T, \delta/\log(T)))^{\frac{1}{3}} \right) \end{aligned}$$

(Here we assume $\overline{\text{Reg}}_{\text{Off}}^H(\mathcal{Q}, t, \delta)$ do not depend on t . Generally this is the case.)

And for locally observable games,

$$V_i = D \cdot \frac{m^2 k^3}{\gamma_i} \quad (\text{for some constant } D)$$

$$\widetilde{\text{Reg}}_T \leq L \cdot \tau_1 + 2L^2 \max_{\ell \in [L]} \frac{\tau_\ell}{2\gamma_\ell} \left(\gamma_1 + \max_{i \in [\ell]} \frac{\gamma_i}{\tau_i} \left(2\tau_i D \cdot m^2 k^3 \frac{1}{\gamma_i} + 18\gamma_i \overline{\text{Reg}}_{\text{Off}}^H(\mathcal{Q}, \tau_i/2, \delta_i) \right) \right).$$

$$\text{set } \gamma_i = \sqrt{\frac{D \cdot \tau_i \cdot k^3 m^2}{\overline{\text{Reg}}_{\text{Off}}^H(\mathcal{Q}, \tau_i/2, \delta_i)}}$$

$$\begin{aligned} \widetilde{\text{Reg}}_T &\leq L \cdot \tau_1 + 2L^2 \max_{\ell \in [L]} \frac{\tau_\ell}{2\gamma_\ell} \left(\gamma_1 + 20 \max_{i \in [\ell]} \frac{\gamma_i}{\tau_i} \left(\sqrt{D \cdot \tau_i \cdot k^3 m^2 \overline{\text{Reg}}_{\text{Off}}^H(\mathcal{Q}, \tau_i/2, \delta_i)} \right) \right) \\ &= L \cdot \tau_1 + 2L^2 \max_{\ell \in [L]} \frac{\tau_\ell}{2\gamma_\ell} (\gamma_1 + 20D \cdot k^3 m^2) \\ &= L \cdot \tau_1 + 2L^2 \max_{\ell \in [L]} \frac{\sqrt{\tau_\ell \overline{\text{Reg}}_{\text{Off}}^H(\mathcal{Q}, \tau_\ell/2, \delta_\ell)}}{\sqrt{D \cdot k^3 m^2}} (\gamma_1 + 20D \cdot k^3 m^2) \end{aligned}$$

$$\begin{aligned}
&= O\left(L^2 \sqrt{T \cdot k^3 m^2 \overline{\text{Reg}}_{\text{Off}}^{\text{H}}(\mathcal{Q}, T/2, \delta/L)}\right) \\
&= O\left(\log^2(T) \sqrt{T \cdot k^3 m^2 \overline{\text{Reg}}_{\text{Off}}^{\text{H}}(\mathcal{Q}, T/2, \delta/\log(T))}\right)
\end{aligned}$$

□

E EMPIRICAL STUDIES

Additional LLM-routing details. The cached BoolQ dataset contains 12,697 prompts, of which 12,648 have parseable responses from both Qwen2.5-0.5B-Instruct, used as the weak model, and Llama-2-13B-Chat, used as the strong reference model. The binary routing outcome indicates whether their cached responses disagree; BoolQ gold labels are not used as routing outcomes. Each prompt is represented by a 138-dimensional context consisting of 10 weak-model uncertainty and probability statistics, 64 whitened principal components of the weak model’s hidden states, and 64 whitened principal components of a context-free prompt embedding. All 12,648 eligible examples are presented online, and outcomes are revealed only after querying the strong model. Each policy makes a routing decision at every round, whereas its estimator is updated only once in 32 rounds, using all feedback revealed before that round. CBPSIDE [Bartók et al., 2012a; An et al., 2026] uses a linear logistic disagreement model and the confidence radius

$$\beta_t(x_t) = \min\left\{0.5s\sqrt{x_t^\top V_{t-1}^{-1}x_t}, 0.5\right\},$$

where s is the tuned multiplier. The two SQUARECB.PMSIDE variants use $\mu = 2$, $\kappa = s\sqrt{T}$, and inverse-propensity weights capped at 10. Their only intended difference is the disagreement-probability estimator. The linear variant fits an L_2 -regularized logistic model with $C = 1$, the 1bfgs solver, and a weighted standardization step. The nonlinear variant uses scikit-learn’s histogram gradient-boosting (HGB) classifier with binary log loss, 50 boosting iterations, learning rate 0.05, at most 15 leaves per tree, a minimum of 20 samples per leaf, L_2 regularization 1, and no early stopping. At each eligible update, both estimators are refitted using their complete accumulated inverse-propensity-weighted feedback. PG-TS [Grant & Leslie, 2021] uses a Bayesian logistic disagreement model with

$$\theta \sim \mathcal{N}(0, (s\sigma_0)^2 I), \quad \sigma_0 = 1,$$

and performs 15 Pólya–Gamma Gibbs transitions whenever its posterior is updated.

Both ETC baselines [Garivier et al., 2016] use an initial exploration budget of $s \cdot T^{2/3}$ rounds to fit a linear-logistic model and a histogram gradient boosting (HGB) model respectively. The fitted models are then frozen, and the remaining queries are routed according to them. The random baseline instead routes each query to the strong model with a fixed probability, independent of context, swept over $\{0, 0.2, 0.4, 0.6, 0.8, 1\}$.

Table 2 shows how s is used to tune different hyper-parameters for different algorithms. For each policy and each ℓ_{01} , we select $s \in \{0.1, 0.3, 1, 3, 10\}$ by minimizing mean total cost over 20 paired random permutations of the online arrival order. Random is computed analytically by matching the routing rate of the selected nonlinear SQUARECB.PMSIDE policy. Figure 2a reports mean realized total cost with pointwise 95% Student- t confidence intervals. The learning curves in Figure 4 report cumulative excess cost relative to a fixed five-fold cross-fitted HGB reference policy. These extend the results discussed in Section 4. Each reference probability is produced by an HGB model that excludes the evaluated example and is used only for evaluation, never as online feedback. Because multiplier selection and evaluation use the same permutations, all selected-policy results are interpreted as exploratory tuned comparisons.

Observations From Figure 2a, it is evident that, both the linear-logistic and nonlinear HGB instantiations of SQUARECB.PMSIDE achieve lower mean realized total cost than CBPSIDE, while the nonlinear HGB variant also outperforms PG-TS. These results illustrate a practical advantage of the oracle-efficient SQUARECB.PMSIDE formulation: it can incorporate flexible nonlinear prediction models without modifying the surrounding routing policy. Furthermore, generally ETC is expected to perform worse due to $O(T^{2/3})$ regret, however here due to limited test data size 12,648 compared to larger dimension 138, it performs competitively. It is evident in the learning curves as well (Figure 4).

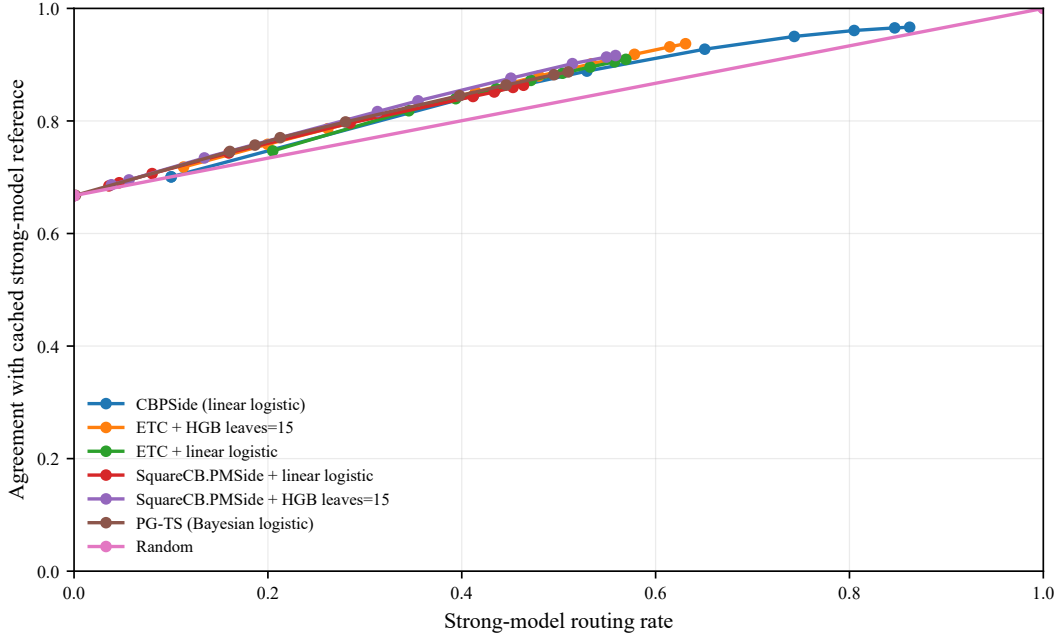


Figure 3: Routing-accuracy trade-off on BoolQ. The horizontal axis reports the fraction of prompts routed to the strong model, while the vertical axis reports agreement with the cached strong-model response. Thus, points closer to the upper-left corner are preferable, as they achieve higher agreement while making fewer strong-model queries. Each marker is the mean over 20 shuffled online orders, with confidence intervals omitted for visual clarity. Along each curve, the connected markers follow increasing values $\ell_{01} \in \{1.8, 2.0, 2.2, 2.4, 2.6, 2.8, 3.0, 3.2, 3.3\}$. Increasing ℓ_{01} assigns a larger penalty to an incorrect weak-model response and therefore generally moves the policy toward more strong-model routing and higher agreement.

Moreover, Figure 3 shows the trade-off between routing accuracy and the rate at which the strong model is invoked; algorithms achieving higher accuracy at a lower invocation rate are preferable. SQUARECB.PMSIDE with the HGB model attains the best trade-off among the algorithms compared.

Table 2: Meaning of the multiplier s for each algorithm.

Algorithm	Scaled quantity	Effect of larger s
CBPSIDE	confidence width sw_i	more exploration
RANCBPSIDE	sampled confidence width sw_i	more exploration
SQUARECB.PMSIDE	IGW rate $\kappa = s\sqrt{T}$	less exploration
ETC	exploration length $sT^{2/3}$	more exploration
PG-TS	prior variance $s\sigma_0$	more exploration

Additional contextual label-efficient prediction experimental details. Contexts are drawn i.i.d. as $x_t \sim \text{Unif}([0, 1]^{10})$, and a single weight vector w is sampled once from $\text{Unif}(0, 1)^{10}$ and normalized so that $w_i \geq 0$ and $\sum_i w_i = 1$; consequently $w^\top x$ is a convex combination of the coordinates of x and lies in $[0, 1]$ by construction. The binary outcome is the deterministic threshold $y_t = \mathbb{1}\{w^\top x_t \geq 0.5\}$. We evaluate $c \in \{0.1, 0.2, 0.3, 0.4, 0.45\}$; we stop short of $c = 0.5$ because at that point the cell of the expert action collapses to a single point: the action becomes degenerate, the two prediction actions become neighbors, and the game falls outside the non-degenerate class that our analysis covers. Each experiment uses horizon $T = 3000$ and is repeated over 20 independent seeds, and for each policy and each c we select $s \in \{0.1, 0.3, 1, 3, 10\}$ by minimizing mean total cost over those 20 paired seeds. Each method has exactly one exploration parameter. To place them on a common footing we scale each by a shared multiplier s , sweeping $s \in \{0.1, 0.3, 1, 3, 10\}$ as summarised in Table 2. Furthermore, The random baseline instead routes each query to the strong model with a fixed probability, independent of context, swept over $\{0, 0.2, 0.4, 0.6, 0.8, 1\}$.

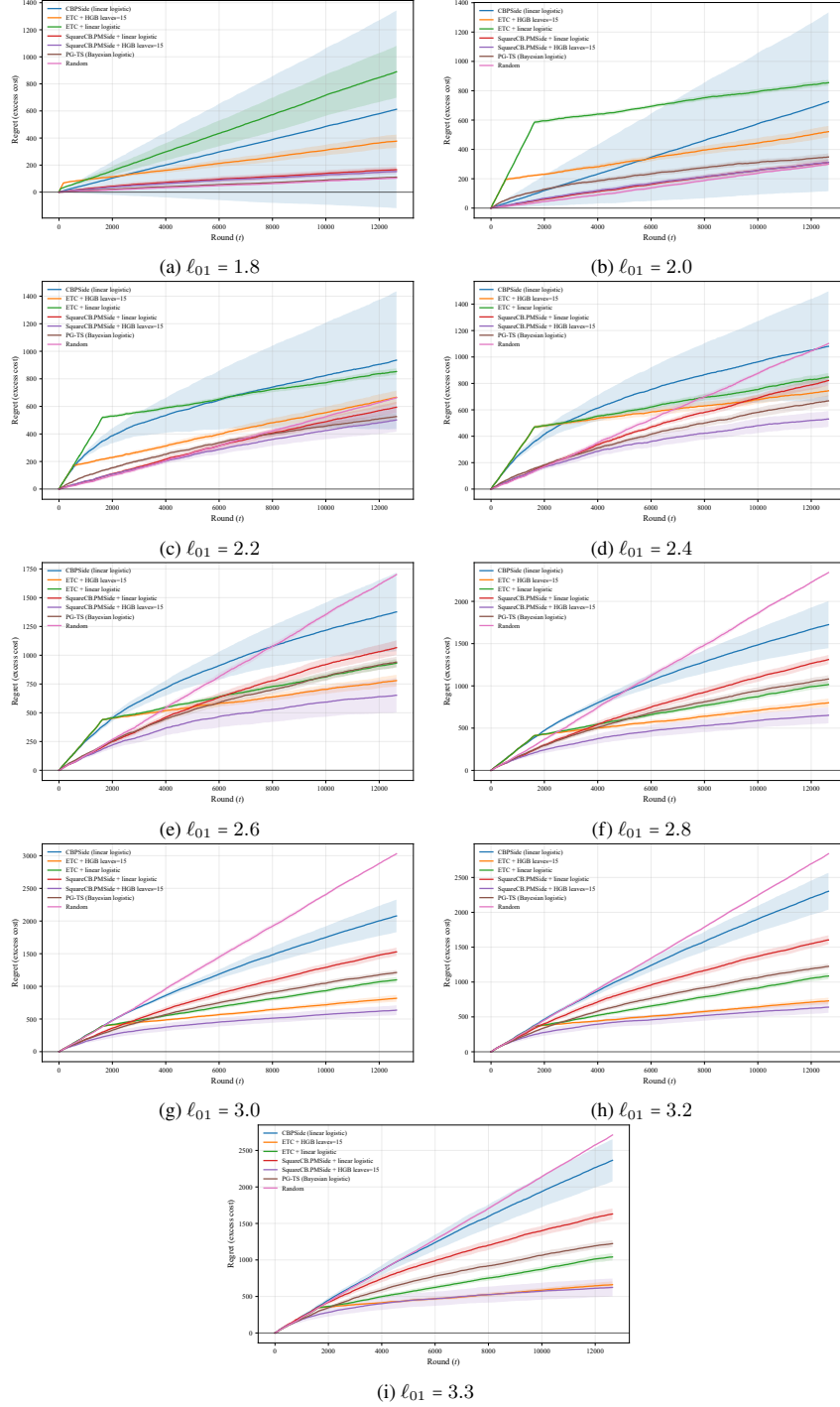


Figure 4: Cumulative empirical regret, measured as excess realized cost relative to the fixed five-fold cross-fitted HGB reference policy, for each value of ℓ_{01} . Curves show means over 20 paired random arrival orders, and shaded regions are pointwise 95% Student- t confidence intervals. Lower values are better; negative regret indicates that the online policy incurs lower realized cost than the fixed reference policy.

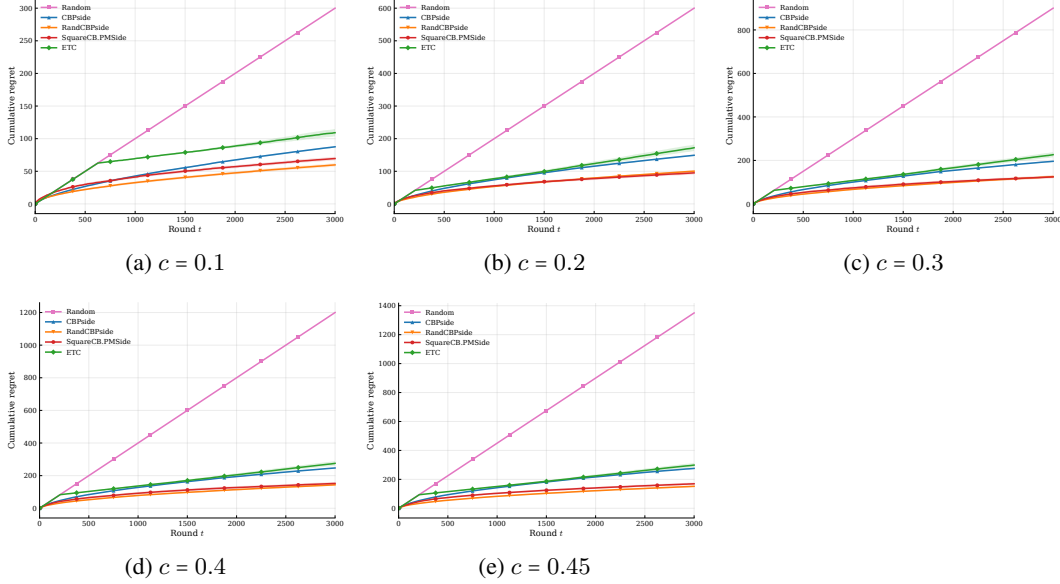


Figure 5: Cumulative regret against round t for each value of the information cost c , with every algorithm at its best multiplier. Shaded bands are 95% Student- t confidence intervals for the mean over 20 seeds.

Observations Figure 2b shows total cost as a function of c , and Figure 5 the learning curves for all compared algorithms across the full sweep. These extend the results discussed in Section 4.

F ALGORITHM FOR Lq -DECREASING TREE CONSTRUCTION

Given a locally observable game $\mathcal{G} = (\mathbf{L}, \Phi)$ and an outcome distribution $q \in \Delta(d)$, we denote $\ell_a := e_a^\top \mathbf{L}q$ for the loss of action a under the outcome distribution q . Since a partial monitoring game can be represented as a neighborhood graph $(\Pi, \mathcal{E})$, we can construct a spanning in-tree $\mathcal{T} \subseteq \mathcal{E}$ rooted at the action with the smallest loss $\hat{a} \in \arg \min_{a \in \Pi} \ell_a$. The following algorithm constructs an Lq -decreasing in-tree by selecting each action's best neighbor to be its parent, then constructing the in-tree $\mathcal{T}$ with the selected neighbors as edges.

Algorithm 5 Constructing an Lq -decreasing in-tree by selecting each action's parent

- 1: **Input:** A locally observable game $\mathcal{G} = (\mathbf{L}, \Phi)$, its neighborhood graph $(\Pi, \mathcal{E})$, and $q \in \Delta(d)$.
 - 2: **Output:** A spanning in-tree $\mathcal{T} \subseteq \mathcal{E}$ with root $\text{root}_{\mathcal{T}} = \hat{a}$ such that Lq is $\mathcal{T}$ -decreasing.
 - 3: For each $a \in \Pi$, compute $\ell_a = e_a^\top \mathbf{L}q$, and choose the root $\hat{a} \in \arg \min_{a \in \Pi} \ell_a$.
 - 4: Choose $q^\circ \in \text{ri}(\mathcal{C}_{\hat{a}})$, which is the relative interior of the cell $\mathcal{C}_{\hat{a}}$.
 - 5: Initialize $\mathcal{T} \leftarrow \emptyset$.
 - 6: **for** each $a \in \Pi \setminus \{\hat{a}\}$ **do**
 - 7: Find the best neighbors $B_a \leftarrow \arg \min_{b: (a,b) \in \mathcal{E}} \ell_b$.
 - 8: Choose $\text{par}_{\mathcal{T}}(a) \in \arg \min_{b \in B_a} e_b^\top \mathbf{L}q^\circ$, breaking any remaining ties arbitrarily.
 - 9: $\mathcal{T} \leftarrow \mathcal{T} \cup \{(a, \text{par}_{\mathcal{T}}(a))\}$.
 - 10: **end for**
 - 11: **return** the in-tree $\mathcal{T}$ with $\text{root}_{\mathcal{T}} = \hat{a}$.
-

The proof of such an Lq -decreasing in-tree construction is provided in Lattimore & Szepesvári [2020, Lemma 19].

G AUXILIARY LEMMATA

Lemma G.1 (Pareto optimal comparators suffice). For every $q \in \Delta(d)$,

$$\min_{\lambda \in \Delta(\Pi)} \lambda^\top \mathbf{L}q = \min_{a \in \Pi} e_a^\top \mathbf{L}q = \min_{a \in [k]} e_a^\top \mathbf{L}q = \min_{\lambda \in \Delta(k)} \lambda^\top \mathbf{L}q.$$

Proof. A linear function on a simplex attains its minimum at a vertex, which gives the outer two equalities, so it suffices to show $\bigcup_{a \in \Pi} C_a = \Delta(d)$. The cells cover $\Delta(d)$ by Definition 2.1, and $\bigcup_{a \in \Pi} C_a$ is a finite union of polytopes of dimension at most $d-2$, hence has empty interior relative to the $(d-1)$ -dimensional set $\Delta(d)$. So $\bigcup_{a \in \Pi} C_a$ is dense in $\Delta(d)$; being a finite union of closed sets it is closed, and therefore equals $\Delta(d)$. $\square$

Theorem D.1 (Regret concentration generalized from Qin & Zhang [2026a]). Let Λ be a set of policies, $\mathcal{L}(\cdot | x) : \Lambda \rightarrow [0, 1]$ a value function and $\mathcal{L}^*(\cdot | x)$ and $\widehat{\mathcal{L}}_\ell(\cdot | x)$, $\ell \in [L]$ are value functions measured under the true model and estimated models respectively. $\lambda^* \in \Lambda$ and $\hat{\lambda}_\ell \in \Lambda$ are respectively the optimal policies under respective true and estimated models. Let $C_\ell : \Lambda \times \Lambda \rightarrow \mathbb{R}_+$ be coverage maps, $p_\ell \in \Lambda$ are decisions, $0 < \gamma_1 \leq \gamma_2 \leq \dots \leq \gamma_L$ learning rates, and $V_\ell, E_\ell \geq 0$ numbers such that,

(C1) (certificate) for every $\ell \geq 1$ and every $\lambda \in \Lambda$,

$$\widehat{\mathcal{L}}_\ell(p_\ell | x) - \widehat{\mathcal{L}}_\ell(\lambda | x) + \frac{1}{\gamma_\ell} C_\ell(p_\ell, \lambda) \leq V_\ell.$$

(C2) (off-policy evaluation) for every $\ell \geq 2$ and every $\lambda \in \Lambda$,

$$\left| \widehat{\mathcal{L}}_\ell(\lambda - p_{\ell-1} | x) - \mathcal{L}^*(\lambda - p_{\ell-1} | x) \right| \leq \frac{1}{6\gamma_\ell} C_{\ell-1}(p_{\ell-1}, \lambda) + \frac{3\gamma_\ell}{2} E_\ell.$$

Define $G_1 := 1$ and, for $\ell \geq 2$,

$$\gamma_\ell G_\ell := \gamma_{\ell-1} G_{\ell-1} + \gamma_{\ell-1} V_{\ell-1} + 9\gamma_\ell^2 E_\ell, \quad \text{i.e.} \quad G_\ell = \frac{1}{\gamma_\ell} \left[\gamma_1 + \sum_{i=2}^\ell (\gamma_{i-1} V_{i-1} + 9\gamma_i^2 E_i) \right].$$

Then, for every $\ell \in [L]$ and every $\lambda \in \Lambda$,

$$\text{Reg}(\lambda | x) \leq 2\widehat{\text{Reg}}_\ell(\lambda | x) + G_\ell \quad \text{and} \quad \widehat{\text{Reg}}_\ell(\lambda | x) \leq 2\text{Reg}(\lambda | x) + G_\ell. \quad (13)$$

Consequently $\text{Reg}(p_\ell | x) \leq 2V_\ell + G_\ell$ for every ℓ , and

$$\sum_{\ell=1}^L (\tau_\ell - \tau_{\ell-1}) \text{Reg}(p_\ell | x) \leq \tau_1 + \sum_{\ell=2}^L \frac{\tau_\ell - \tau_{\ell-1}}{\gamma_\ell} \left[2\gamma_\ell V_\ell + \gamma_1 + \sum_{i=2}^\ell (\gamma_{i-1} V_{i-1} + 9\gamma_i^2 E_i) \right]. \quad (14)$$

Proof. We prove Eq. equation 13 by induction on ℓ . For $\ell = 1$ both inequalities hold with $G_1 = 1$, since Reg and $\widehat{\text{Reg}}_1$ take values in $[0, 1]$. Fix $\ell \geq 2$ and assume Eq. equation 13 to hold at $\ell - 1$.

First, we establish that, for every λ , our policy and estimator guarantees Low Regret (LR) and Good Coverage(GC).

That is, from (C1) for epoch $\ell - 1$, we have

$$\begin{aligned} \widehat{\mathcal{L}}_{\ell-1}(p_{\ell-1} | x) - \widehat{\mathcal{L}}_{\ell-1}(\lambda | x) + \frac{1}{\gamma_{\ell-1}} C_{\ell-1}(p_{\ell-1}, \lambda) &\leq V_{\ell-1} \\ \widehat{\mathcal{L}}_{\ell-1}(p_{\ell-1} | x) - \widehat{\mathcal{L}}_{\ell-1}(\hat{\lambda}_{\ell-1} | x) - (\widehat{\mathcal{L}}_{\ell-1}(\lambda | x) - \widehat{\mathcal{L}}_{\ell-1}(\hat{\lambda}_{\ell-1} | x)) + \frac{1}{\gamma_{\ell-1}} C_{\ell-1}(p_{\ell-1}, \lambda) &\leq V_{\ell-1} \\ \widehat{\text{Reg}}_{\ell-1}(p_{\ell-1} | x) + \left(\frac{1}{\gamma_{\ell-1}} C_{\ell-1}(p_{\ell-1}, \lambda) - \widehat{\text{Reg}}_{\ell-1}(\lambda | x) \right) &\leq V_{\ell-1} \end{aligned}$$

Above statement is true for all λ , including the the case where $\widehat{\text{Reg}}_{\ell-1}(\lambda | x) = 0$ and $C_\ell(p_{\ell-1}, \lambda)$ is always ≥ 0 . Hence,

$$\widehat{\text{Reg}}_{\ell-1}(p_{\ell-1} | x) \leq V_{\ell-1}. \quad (\text{Low Regret (LR)})$$

Also,

$$\left(\frac{1}{\gamma_{\ell-1}} C_{\ell-1}(p_{\ell-1}, \lambda) - \widehat{\text{Reg}}_{\ell-1}(\lambda | x) \right) \leq V_{\ell-1}.$$

Hence,

$$C_{\ell-1}(p_{\ell-1}, \lambda) \leq \gamma_{\ell-1} \left[V_{\ell-1} + \widehat{\text{Reg}}_{\ell-1}(\lambda | x) \right]. \quad (\text{Good Coverage (GC)})$$

Moreover,

$$\begin{aligned}
& \text{Reg}(\lambda | x) - \widehat{\text{Reg}}_{\ell}(\lambda | x) \\
&= \mathcal{L}^*(\lambda | x) - \mathcal{L}^*(\lambda^* | x) - \left(\widehat{\mathcal{L}}_{\ell}(\lambda | x) - \widehat{\mathcal{L}}_{\ell}(\hat{\lambda}_{\ell} | x) \right) \\
&= \mathcal{L}^*(\lambda - p_{\ell-1} | x) - \mathcal{L}^*(\lambda^* - p_{\ell-1} | x) - \left(\widehat{\mathcal{L}}_{\ell}(\lambda - p_{\ell-1} | x) - \widehat{\mathcal{L}}_{\ell}(\hat{\lambda}_{\ell} - p_{\ell-1} | x) \right) \\
&= \mathcal{L}^*(\lambda - p_{\ell-1} | x) - \widehat{\mathcal{L}}_{\ell}(\lambda - p_{\ell-1} | x) - \left(\mathcal{L}^*(\lambda^* - p_{\ell-1} | x) - \widehat{\mathcal{L}}_{\ell}(\hat{\lambda}_{\ell} - p_{\ell-1} | x) \right) \\
&\leq \mathcal{L}^*(\lambda - p_{\ell-1} | x) - \widehat{\mathcal{L}}_{\ell}(\lambda - p_{\ell-1} | x) - \left(\mathcal{L}^*(\lambda^* - p_{\ell-1} | x) - \widehat{\mathcal{L}}_{\ell}(\lambda^* - p_{\ell-1} | x) \right) \\
&\hspace{15em} (\widehat{\mathcal{L}}_{\ell}(\lambda^* | x) \geq \widehat{\mathcal{L}}_{\ell}(\hat{\lambda}_{\ell} | x)) \\
&\leq |\mathcal{L}^*(\lambda - p_{\ell-1} | x) - \widehat{\mathcal{L}}_{\ell}(\lambda - p_{\ell-1} | x)| + |\mathcal{L}^*(\lambda^* - p_{\ell-1} | x) - \widehat{\mathcal{L}}_{\ell}(\lambda^* - p_{\ell-1} | x)| \\
&\leq \frac{1}{6\gamma_{\ell}} C_{\ell-1}(p_{\ell-1}, \lambda) + \frac{3\gamma_{\ell}}{2} E_{\ell} + \frac{1}{6\gamma_{\ell}} C_{\ell-1}(p_{\ell-1}, \lambda^*) + \frac{3\gamma_{\ell}}{2} E_{\ell} \quad (\text{C2}) \\
&\leq \frac{1}{6\gamma_{\ell}} \gamma_{\ell-1} \left[V_{\ell-1} + \widehat{\text{Reg}}_{\ell-1}(\lambda | x) \right] + \frac{3\gamma_{\ell}}{2} E_{\ell} + \frac{1}{6\gamma_{\ell}} \gamma_{\ell-1} \left[V_{\ell-1} + \widehat{\text{Reg}}_{\ell-1}(\lambda^* | x) \right] + \frac{3\gamma_{\ell}}{2} E_{\ell} \\
&\hspace{15em} (\text{equation Good Coverage (GC)}) \\
&\leq \frac{1}{6\gamma_{\ell}} \gamma_{\ell-1} \left[V_{\ell-1} + 2\text{Reg}(\lambda | x) + G_{\ell-1} \right] + \frac{3\gamma_{\ell}}{2} E_{\ell} + \frac{1}{6\gamma_{\ell}} \gamma_{\ell-1} \left[V_{\ell-1} + 2\text{Reg}(\lambda^* | x) + G_{\ell-1} \right] + \frac{3\gamma_{\ell}}{2} E_{\ell} \\
&\hspace{15em} (\text{equation 13 true for } \ell - 1) \\
&\leq \frac{1}{6\gamma_{\ell}} \gamma_{\ell-1} \left[D_{\ell-1} + 2\text{Reg}(\lambda | x) + G_{\ell-1} \right] + \frac{3\gamma_{\ell}}{2} E_{\ell} + \frac{1}{6\gamma_{\ell}} \gamma_{\ell-1} \left[V_{\ell-1} + G_{\ell-1} \right] + \frac{3\gamma_{\ell}}{2} E_{\ell} \\
&\hspace{15em} (\text{Reg}(\lambda^* | x) = 0) \\
&= \frac{1}{3\gamma_{\ell}} (\gamma_{\ell-1} G_{\ell-1} + \gamma_{\ell-1} V_{\ell-1} + 9\gamma_{\ell}^2 E_{\ell}) + \frac{\gamma_{\ell-1}}{3\gamma_{\ell}} \text{Reg}(\lambda | x) \\
&\leq \frac{1}{3} \text{Reg}(\lambda | x) + \frac{1}{3} G_{\ell} \quad (\gamma_{\ell-1} \leq \gamma_{\ell})
\end{aligned}$$

Hence

$$\begin{aligned}
\text{Reg}(\lambda | x) - \widehat{\text{Reg}}_{\ell}(\lambda | x) &\leq \frac{1}{3} \text{Reg}(\lambda | x) + \frac{1}{3} G_{\ell} \\
\frac{2}{3} \text{Reg}(\lambda | x) &\leq \widehat{\text{Reg}}_{\ell}(\lambda | x) + \frac{1}{3} G_{\ell} \\
\text{Reg}(\lambda | x) &\leq \frac{3}{2} \widehat{\text{Reg}}_{\ell}(\lambda | x) + \frac{1}{2} G_{\ell} \\
&\leq 2\widehat{\text{Reg}}_{\ell}(\lambda | x) + G_{\ell} \quad (\text{Step **})
\end{aligned}$$

Furthermore,

$$\begin{aligned}
& \widehat{\text{Reg}}_{\ell}(\lambda | x) - \text{Reg}(\lambda | x) \\
&= \widehat{\mathcal{L}}_{\ell}(\lambda | x) - \widehat{\mathcal{L}}_{\ell}(\hat{\lambda}_{\ell} | x) - (\mathcal{L}^*(\lambda | x) - \mathcal{L}^*(\lambda^* | x)) \\
&= \widehat{\mathcal{L}}_{\ell}(\lambda - p_{\ell-1} | x) - \widehat{\mathcal{L}}_{\ell}(\hat{\lambda}_{\ell} - p_{\ell-1} | x) - (\mathcal{L}^*(\lambda - p_{\ell-1} | x) - \mathcal{L}^*(\lambda^* - p_{\ell-1} | x)) \\
&= \widehat{\mathcal{L}}_{\ell}(\lambda - p_{\ell-1} | x) - \mathcal{L}^*(\lambda - p_{\ell-1} | x) - (\widehat{\mathcal{L}}_{\ell}(\hat{\lambda}_{\ell} - p_{\ell-1} | x) - \mathcal{L}^*(\lambda^* - p_{\ell-1} | x)) \\
&\leq \widehat{\mathcal{L}}_{\ell}(\lambda - p_{\ell-1} | x) - \mathcal{L}^*(\lambda - p_{\ell-1} | x) - (\widehat{\mathcal{L}}_{\ell}(\hat{\lambda}_{\ell} - p_{\ell-1} | x) - \mathcal{L}^*(\hat{\lambda}_{\ell} - p_{\ell-1} | x)) \\
&\hspace{15em} (\mathcal{L}^*(\hat{\lambda}_{\ell} | x) \geq \mathcal{L}^*(\lambda^* | x)) \\
&\leq |\widehat{\mathcal{L}}_{\ell}(\lambda - p_{\ell-1} | x) - \mathcal{L}^*(\lambda - p_{\ell-1} | x)| + |\widehat{\mathcal{L}}_{\ell}(\hat{\lambda}_{\ell} - p_{\ell-1} | x) - \mathcal{L}^*(\hat{\lambda}_{\ell} - p_{\ell-1} | x)|
\end{aligned}$$

$$\begin{aligned}
&\leq \frac{1}{6\gamma_\ell} C_{\ell-1}(p_{\ell-1}, \lambda) + \frac{3\gamma_\ell}{2} E_\ell + \frac{1}{6\gamma_\ell} C_{\ell-1}(p_{\ell-1}, \hat{\lambda}_\ell) + \frac{3\gamma_\ell}{2} E_\ell \quad (\text{C2}) \\
&\leq \frac{1}{6\gamma_\ell} \gamma_{\ell-1} [V_{\ell-1} + \widehat{\text{Reg}}_{\ell-1}(\lambda | x)] + \frac{3\gamma_\ell}{2} E_\ell + \frac{1}{6\gamma_\ell} \gamma_{\ell-1} [V_{\ell-1} + \widehat{\text{Reg}}_{\ell-1}(\hat{\lambda}_\ell | x)] + \frac{3\gamma_\ell}{2} E_\ell \\
&\quad (\text{equation Good Coverage (GC)}) \\
&\leq \frac{1}{6\gamma_\ell} \gamma_{\ell-1} [V_{\ell-1} + 2\text{Reg}(\lambda | x) + G_{\ell-1}] + \frac{3\gamma_\ell}{2} E_\ell + \frac{1}{6\gamma_\ell} \gamma_{\ell-1} [V_{\ell-1} + 2\text{Reg}(\hat{\lambda}_\ell | x) + G_{\ell-1}] + \frac{3\gamma_\ell}{2} E_\ell \\
&\quad (\text{equation 13 true for } \ell - 1) \\
&\leq \frac{1}{6\gamma_\ell} \gamma_{\ell-1} [V_{\ell-1} + 2\text{Reg}(\lambda | x) + G_{\ell-1}] + \frac{3\gamma_\ell}{2} E_\ell + \frac{1}{6\gamma_\ell} \gamma_{\ell-1} [V_{\ell-1} + 4\widehat{\text{Reg}}_\ell(\hat{\lambda}_\ell | x) + 2G_\ell + G_{\ell-1}] \\
&\quad + \frac{3\gamma_\ell}{2} E_\ell \quad (\text{equation 13 true for } \ell \text{ (Step **)}) \\
&= \frac{1}{6\gamma_\ell} \gamma_{\ell-1} [V_{\ell-1} + 2\text{Reg}(\lambda | x) + G_{\ell-1}] + \frac{3\gamma_\ell}{2} E_\ell + \frac{1}{6\gamma_\ell} \gamma_{\ell-1} [V_{\ell-1} + 2G_\ell + G_{\ell-1}] + \frac{3\gamma_\ell}{2} E_\ell \\
&\quad (\widehat{\text{Reg}}_\ell(\hat{\lambda}_\ell | x) = 0) \\
&= \frac{1}{3\gamma_\ell} (\gamma_{\ell-1} G_{\ell-1} + \gamma_{\ell-1} V_{\ell-1} + 9\gamma_\ell^2 E_\ell) + \frac{\gamma_{\ell-1}}{3\gamma_\ell} \text{Reg}(\lambda | x) + \frac{\gamma_{\ell-1}}{3\gamma_\ell} G_\ell \\
&\leq \frac{1}{3} \text{Reg}(\lambda | x) + \frac{2}{3} G_\ell \quad (\gamma_{\ell-1} \leq \gamma_\ell)
\end{aligned}$$

Hence,

$$\begin{aligned}
\widehat{\text{Reg}}_\ell(\lambda | x) - \text{Reg}(\lambda | x) &\leq \frac{1}{3} \text{Reg}(\lambda | x) + \frac{2}{3} G_\ell \\
\widehat{\text{Reg}}_\ell(\lambda | x) &\leq \frac{4}{3} \text{Reg}(\lambda | x) + \frac{2}{3} G_\ell \\
\text{Reg}(\lambda | x) &\leq \frac{3}{2} \widehat{\text{Reg}}_\ell(\lambda | x) + \frac{1}{2} G_\ell \\
&\leq 2\widehat{\text{Reg}}_\ell(\lambda | x) + G_\ell
\end{aligned}$$

Finally, we prove the consequences. Apply (C1) at $\lambda = \hat{\lambda}_\ell$, we have

$$\begin{aligned}
\widehat{\text{Reg}}_\ell(p_\ell | x) &\leq V_\ell \quad (\text{equation Low Regret (LR)}) \\
\Rightarrow \text{Reg}(p_\ell | x) &\leq 2\widehat{\text{Reg}}_\ell(p_\ell | x) + G_\ell \leq 2V_\ell + G_\ell. \quad (\text{By equation 13})
\end{aligned}$$

Summing over epochs, we have

$$\begin{aligned}
\sum_{\ell=1}^L (\tau_\ell - \tau_{\ell-1}) \text{Reg}(p_\ell | x) &\leq \tau_1 + \sum_{\ell=2}^L (\tau_\ell - \tau_{\ell-1}) (2V_\ell + G_\ell) \quad (\text{Reg}(p_1 | x) \leq 1) \\
&= \tau_1 + \sum_{\ell=2}^L \frac{\tau_\ell - \tau_{\ell-1}}{\gamma_\ell} \left[2\gamma_\ell V_\ell + \gamma_1 + \sum_{i=2}^{\ell} (\gamma_{i-1} V_{i-1} + 9\gamma_i^2 E_i) \right], \\
&\quad (\text{By the closed form of } G_\ell)
\end{aligned}$$

which is Eq. equation 14. $\square$

Lemma G.2. Let P, Q are two probability simplices. $D_H(P|Q)$, $D_{KL}(P|Q)$, $\|P-Q\|_{TV}$, $\|P-Q\|_1$ be the *Hellinger distance*, *Kullback divergence*, *total variation distance* and *level-1 norm* between P, Q respectively, then the following holds,

$$\frac{\|P-Q\|_1}{2} = \|P-Q\|_{TV} \leq D_H(P|Q)$$

and,

$$\frac{\|P-Q\|_1}{2} = \|P-Q\|_{TV} \leq \sqrt{D_{KL}(P|Q)}$$

Proof.

$$\|P - Q\|_{\text{TV}} \leq D_{\text{H}}(P \mid Q) \quad (\text{Tsybakov [2008], Lemma 2.3})$$

and,

$$\|P - Q\|_{\text{TV}} \leq \sqrt{D_{\text{KL}}(P \mid Q)} \quad (\text{Tsybakov [2008], Lemma 2.5})$$

Furthermore,

$$\frac{\|P - Q\|_1}{2} = \|P - Q\|_{\text{TV}} \quad (\text{Levin \& Peres [2017], Proposition 4.2})$$

□

Lemma G.3 (Local observability, Lattimore & Szepesvári [2020]). If $\mathcal{G} = (\Phi, \mathbf{L})$ is locally observable and non-degenerate and actions $e = (a, b) \in \mathcal{E}$ are neighbours, then there exist functions $w_e(a, \cdot), w_e(b, \cdot) : \Sigma \rightarrow \mathbb{R}$ such that $\|w_e\|_{\infty} \leq m/2$ and

$$\mathbf{L}_{ao} - \mathbf{L}_{bo} = w_e(a, \Phi_{ao}) + w_e(b, \Phi_{bo}) \text{ for all } o \in [d].$$

In Lemma G.3, the assumption of non-degeneracy is important here for $\|w_e\|_{\infty} \leq m/2$ to hold. The Proposition G.4 bounds $\|w_e\|_{\infty}$ for more general case.

Lemma G.4 (Lattimore & Szepesvári [2020]). Suppose that (L, Φ) is globally observable and a and b are neighbours. Then there exists a function $w_e : [k] \times \Sigma \rightarrow \mathbb{R}$ such that for all $x \in [d]$,

$$\sum_{c=1}^k w_e(c, \Phi_{co}) = L_{ao} - L_{bo}.$$

Furthermore, w_e can be chosen so that $\|w_e\|_{\infty} \leq d^{1/2} k^{d/2}$.

Lemma G.5 (Water-transfer operator adapted from Lattimore & Szepesvári [2020]). Suppose that $\mathcal{G}$ is non-degenerate and locally observable and $q \in \Delta(d)$. Assume that we have an in-tree $\mathcal{T}$ that is Lq -decreasing. Then there exists a function $W_q^{\mathcal{T}} : \Delta(\Pi) \rightarrow \Delta(\Pi)$ such that the following hold for all $p \in \Delta(\Pi)$:

- (a) The expected loss does not increase: $(W_q(p) - p)^{\top} \mathbf{L}q \leq 0$.
- (b) Action probabilities are not too small: $W_q(p)[a] \geq p[a]/k$ for all $a \in [k]$.
- (c) Probabilities increase towards the root of the in-tree $\mathcal{T}$. That is, $W_q(p)[a] \leq W_q(p)[b]$ for all $e = (a, b) \in \mathcal{T}$.